

APPRENDRE, TOUJOURS

PEARSON

Le présent ouvrage est la traduction de *Biostatistics for the Biological and Health Sciences* de Marc M. Triola, M.D. et Mario F. Triola, publié par Pearson Education Inc., Copyright © 2006 par Pearson Education, Inc.

Publié par Pearson France
Immeuble Terra Nova II
15 rue Henri Rol-Tanguy
93100 MONTREUIL
Tél. : +33(0)1 43 62 31 00
www.pearson.fr

Mise en pages : TyPAO

ISBN : 978-2-7440-7657-2

Table des matières

Table des symboles

$\bar{A}$	complémentaire de l'événement A
H_0	hypothèse nulle
H_1	hypothèse alternative
α	(alpha) probabilité d'une erreur de première espèce ou aire de la région critique
β	(bêta) probabilité d'une erreur de deuxième espèce
r	coefficient de corrélation linéaire de l'échantillon
ρ	(rhô) coefficient de corrélation linéaire de la population
r^2	coefficient de détermination
R^2	coefficient de détermination multiple
r_s	coefficient de corrélation des rangs de Spearman
b_1	estimation ponctuelle de la pente de la droite de régression
b_0	estimation ponctuelle de l'ordonnée à l'origine pour la droite de régression
$\hat{y}$	valeur prédite de y
d	différence entre deux valeurs appariées
$\bar{d}$	moyenne des différences d entre les valeurs appariées d'échantillon
s_d	écart type des différences d entre les valeurs appariées d'échantillon
s_e	erreur standard de l'estimation
T	somme des rangs ; utilisée dans le test des rangs signés de Wilcoxon
H	statistique de test de Kruskal-Wallis
R	somme des rangs pour un échantillon ; utilisée dans le test de la somme des rangs de Wilcoxon
μ_R	moyenne attendue des rangs ; utilisée dans le test de la somme des rangs de Wilcoxon
σ_R	écart type attendu des rangs ; utilisé dans le test de la somme des rangs de Wilcoxon
$\mu_{\bar{x}}$	moyenne de la population de toutes les moyennes d'échantillon possibles $\bar{x}$
$\sigma_{\bar{x}}$	écart type de la population de toutes les moyennes d'échantillon possibles $\bar{x}$
E	marge d'erreur de l'estimation du paramètre de la population ou valeur attendue
Q_1, Q_2, Q_3	quartiles
$D_1, D_2 \dots D_9$	déciles
$P_1, P_2 \dots P_{99}$	percentiles
x	valeur d'une donnée
f	fréquence d'apparition d'une valeur
Σ	(sigma majuscule) sommation
Σx	somme des valeurs
Σx^2	somme des carrés des valeurs

1 Introduction

Problème du chapitre 1

Que pouvons-nous apprendre de cette enquête de santé ?

USA Today a réalisé une enquête de santé qui remplissait 3/4 de page dans un de ses numéros. On demandait aux lecteurs de « prendre un moment pour remplir et renvoyer le formulaire ». Les lecteurs pouvaient envoyer leurs réponses par courrier électronique ou par fax. La première question demandait combien de fois ils voyaient un médecin par an. La seconde les interrogeait sur un bilan de santé pour l'année passée incluant grippe, fièvre, hémorroïdes et verrues. La plupart des questions traitaient de conditions de santé, d'usage du tabac et de médicaments. La question 17 était : « Pouvons-nous vous contacter pour participer à d'autres enquêtes de *USA Today* ? ». Les lecteurs qui y répondaient positivement devaient alors fournir leur adresse, leur(s) numéro(s) de téléphone et leur adresse-mail.

Considérons la façon dont les données sont collectées dans cette enquête. En quoi cela affecte-t-il nos conclusions quant à la population générale si on se base sur les résultats obtenus à partir de ce genre d'enquête ? Pouvons-nous utiliser les nombres de visites chez les médecins fournis pour estimer le nombre de visites dans la population générale ? Les réponses à de telles questions sont vitales pour l'évaluation des résultats de telles enquêtes.

Le sujet qui est abordé ici est le point le plus important de tout ce chapitre et ce pourrait bien être le point le plus important de l'ensemble de ce livre.

Dans ce chapitre nous allons nous intéresser à la validité de telles enquêtes. Nous verrons que nous pouvons souvent tirer des conclusions importantes à partir de simples règles de bon sens. Après avoir lu ce chapitre, vous devriez être capables d'identifier les points clés qui affectent la validité de l'enquête précédente et vous devriez avoir une bonne compréhension des méthodes de collecte des données en général.

L'état des statistiques

Le mot *statistiques* est dérivé du mot latin *status* (qui signifie « état »). Des usages très précoces des statistiques se retrouvent dans la compilation de données et de graphiques décrivant divers aspects d'un pays ou d'une région. En 1662, John Graunt a publié des informations statistiques sur les naissances et les décès. Le travail de Graunt fut suivi par des études sur la mortalité, les taux de maladies, les tailles de populations, les revenus et les taux de chômage. Les foyers, les gouvernements et le monde du travail s'appuient fortement sur les statistiques pour se guider. Par exemple, les taux de chômage ou d'inflation, les indices de consommation sont soigneusement compilés de façon régulière et les données qui en résultent sont utilisées par les chefs d'entreprise pour prendre des décisions qui affectent les achats futurs, les niveaux de production et l'expansion vers de nouveaux marchés.

Devriez-vous croire à une étude statistique ?

Dans la seconde édition de *Statistical Reasoning for Everyday Life*, les auteurs Jeff Bennett, William Briggs et Mario Triola listent les 8 points fondamentaux pour évaluer de façon critique une étude statistique : (1) identifier le but de l'étude, la population considérée et le type d'étude ; (2) considérer les sources, en particulier au regard d'une possibilité de biais ; (3) analyser la méthode d'échantillonnage ; (4) chercher les problèmes de définition ou de mesure des variables d'intérêt ; (5) surveiller les variables confondantes qui pourraient invalider les conclusions ; (6) considérer le cadre et la formulation de l'enquête ; (7) vérifier que les graphiques représentent fidèlement les données et que les conclusions sont justifiées ; (8) déterminer si les conclusions répondent au but de l'enquête, si elles ont du sens et si elles ont une signification pratique.



Le **niveau intervalle de mesure** est semblable au niveau ordinal avec la propriété supplémentaire que la différence entre deux valeurs a un sens. Cependant, à ce niveau, les données n'ont pas de zéro *naturel* de référence (pour lequel *aucune* quantité n'est présente).

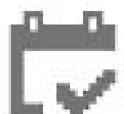
Exemples

1. Températures : les températures du corps humain comme 36,8 °C et 37,0 °C sont des exemples de données au niveau intervalle. Ces valeurs sont ordonnées et nous pouvons déterminer que leur différence est de 0,2 °C. Cependant il n'y a pas de zéro naturel de référence. La valeur de 0 °C pourrait sembler être un point de référence mais c'est une valeur arbitraire et cela ne représente pas l'absence totale de chaleur. Parce que 0 °C n'est pas un zéro naturel de référence, il est faux de dire que 50 °C est *deux fois* plus chaud que 25 °C.

2. Années d'apparition des cigales : les années 1936, 1953, 1970, 1987 et 2004 (le temps n'a pas commencé à l'année 0, ainsi l'année 0 est arbitraire au lieu d'être un zéro naturel de référence représentant « pas de temps »).



Le **niveau rapport de mesure** est semblable au niveau intervalle avec la propriété supplémentaire qu'il y a un zéro naturel de référence pour lequel *aucune* quantité n'est présente. Pour les valeurs à ce niveau, les *différences* et les *rapports* ont un sens.



Exemples

On notera l'utilisation des rapports « deux fois » et « trois fois ».

1. Poids : les poids (en kg) des aigles (0 kg représente l'absence de poids et 4 kg est deux fois plus lourd que 2 kg).

2. Âges : les âges (en jours) des aigles (0 représente un nouveau-né sans âge et un aigle de 60 jours est trois fois plus vieux qu'un aigle de 20 jours).

Ce niveau de mesure est appelé le niveau rapport parce que la valeur 0 de référence donne un sens aux rapports de valeurs. Parmi les 4 niveaux de mesure, la plus grande difficulté est de distinguer les niveaux intervalle et rapport. *Indication* : pour faciliter cette distinction, utilisez un simple

Dans un de ses livres, David Salsburg cite le cas d'une étude rétrospective qui montrait que des édulcorants artificiels étaient liés au cancer de la vessie. Cependant la plupart des sujets malades venaient des classes économiquement faibles alors que la plupart des sujets non malades venaient des classes économiquement supérieures. En conséquence, les deux groupes n'étaient pas comparables et cette étude rétrospective était faussée.

Dans les études prospectives, nous avançons dans le temps en suivant des groupes soumis à des effets d'un facteur potentiel et d'autres, non soumis à de tels effets, comme un groupe de conducteurs qui utilisent des téléphones portables et un groupe qui n'en utilise pas.

Les trois définitions précédentes s'appliquent aux études observationnelles, mais nous allons maintenant nous intéresser aux études expérimentales. Les résultats des expériences sont parfois faussées à cause de la *confusion*.



La **confusion** survient quand des effets de variables sont mélangés et que les effets *individuels* des variables ne peuvent pas être identifiés (c'est-à-dire que la confusion est fondamentalement la confusion des effets des variables).

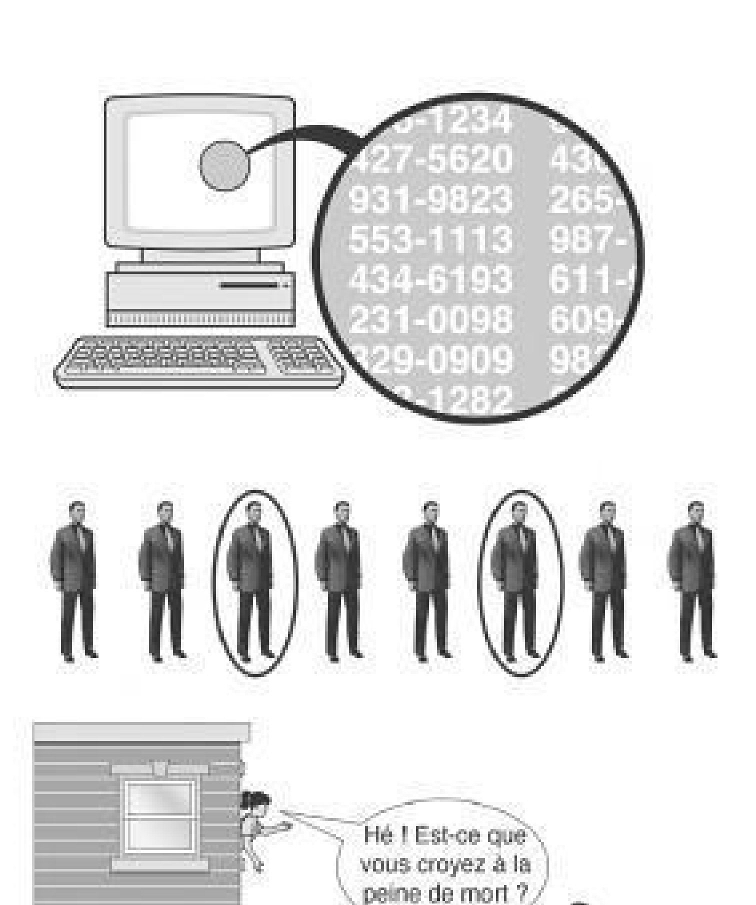
Essayez d'organiser vos expériences afin d'éviter la confusion.

Par exemple, supposons que nous traitions 1 000 personnes avec un vaccin prévu pour prévenir la maladie de Lyme causée par les tiques. Si un froid précoce fait hiberner les tiques et que les 1 000 sujets vaccinés montrent en conséquence une faible incidence de la maladie de Lyme, nous ne pouvons pas savoir si la baisse du taux de maladie est le résultat de l'action du vaccin ou de la survenue précoce du froid. La confusion est apparue parce que les effets du traitement par le vaccin et les effets du froid sont mélangés. Une meilleure planification expérimentale devrait mieux prendre en compte l'effet du vaccin et l'action du froid pour que leurs effets respectifs soient identifiés et contrôlés.

Contrôler les effets des variables

La figure 1.1 montre qu'un des éléments clés dans la conception des expériences est de contrôler les effets des variables. On peut obtenir ce contrôle avec des techniques comme l'insu, les blocs, une étude complètement randomisée, ou une étude expérimentale rigoureusement contrôlée dont la description suit.

Insu En 1954, une étude de masse fut organisée pour tester l'efficacité du vaccin Salk pour prévenir la polio qui paralysait ou tuait des milliers d'enfants. Dans cette étude, un groupe traitement reçut le vaccin Salk alors qu'un second groupe recevait un placebo qui ne contenait aucun vaccin. Dans les études mettant en jeu les placebos, il y a souvent un **effet placebo** qui apparaît quand un sujet montre une amélioration des symptômes (l'amélioration rapportée dans le groupe placebo peut être réelle ou imaginée). Cet effet placebo peut être minimisé ou comptabilisé à travers la technique d'**insu** (ou d'*aveugle*), technique où le sujet ne sait pas s'il reçoit un traitement ou un placebo. L'insu nous permet de déterminer si l'effet du traitement est significativement différent de l'effet placebo. Dans une étude **simple aveugle**, les sujets ne savent pas s'ils reçoivent un traitement ou un placebo. L'étude polio était en **double aveugle**, ce qui signifie qu'il y avait deux niveaux d'aveuglement : (1) les enfants ne savaient pas s'ils recevaient le vaccin Salk ou un placebo et (2) les médecins qui faisaient les injections et évaluaient les résultats ne le savaient pas eux-mêmes.

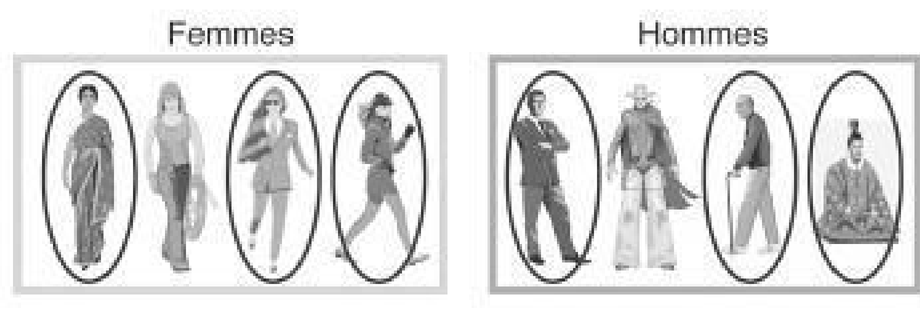


Échantillonnage aléatoire :
Chaque membre de la population a la même chance d'être choisi. Les ordinateurs sont souvent utilisés pour générer des nombres aléatoires.

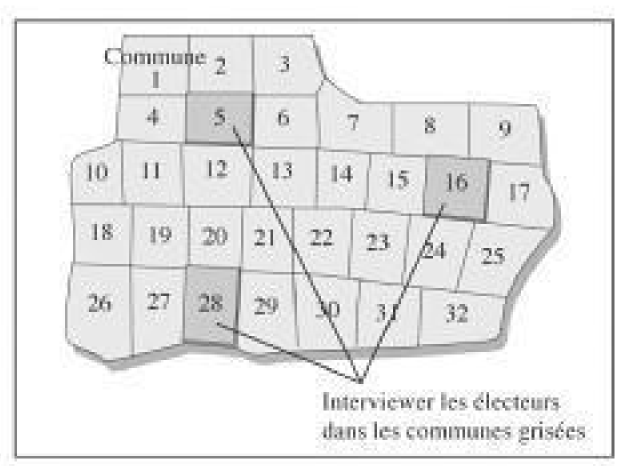
Échantillonnage aléatoire simple :
Un échantillon aléatoire simple de n sujets est choisi de telle façon que chaque échantillon possible de taille n ait la même chance d'être choisi.

Échantillonnage systématique :
Prenez un point de départ, puis sélectionnez chaque k-ième (par exemple chaque $50^{\text{ième}}$) élément de la population.

Échantillonnage opportun :
Utilisez les résultats qui sont faciles à obtenir.



Échantillonnage stratifié :
Subdivisez la population en au moins deux sous-groupes différents (ou strates) qui partagent les mêmes caractéristiques (comme le sexe, la classe d'âge) puis tirez un échantillon dans chaque sous-groupe.



Échantillonnage en grappes :
Divisez la population en sections (ou grappes), puis sélectionnez aléatoirement des grappes et choisissez tous les membres des grappes sélectionnées.

Figure 1.2 Méthodes usuelles d'échantillonnage



Une erreur d'échantillonnage est la différence entre un résultat d'échantillonnage et le vrai résultat de la population ; une telle erreur résulte des fluctuations de tirage de l'échantillon.

Une **erreur non liée à l'échantillonnage** survient lorsque les données d'échantillon sont incorrectement collectées, enregistrées ou analysées (comme en sélectionnant un échantillon biaisé, en utilisant un instrument de mesure défectueux ou en enregistrant incorrectement les données).



Une **distribution de fréquences** liste les valeurs des données (soit individuellement soit par intervalles) et les fréquences correspondantes (ou comptages).

Le tableau 2-2 est une distribution de fréquences résumant les niveaux mesurés de cotinine des 40 fumeurs listés dans le tableau 2-1. La **fréquence** pour une classe donnée est le nombre de données originales qui sont dans cette classe. Par exemple, la première classe du tableau 2-2 a une fréquence de 11, ce qui signifie que 11 des valeurs originales sont entre 0 et 99 inclus.

Nous allons d'abord présenter quelques termes standard utilisés dans la discussion des distributions de fréquences et nous décrirons ensuite comment les construire et les interpréter.

Tableau 2-2 Distribution de fréquences des niveaux de cotinine des fumeurs

Niveau de cotinine (ng/ml)	Fréquence (nombre de fumeurs)
0-99	11
100-199	12
200-299	14
300-399	1
400-499	2



Les **limites de classe inférieures** sont les plus petits nombres qui peuvent appartenir aux différentes classes. Le tableau 2-2 a comme limites de classe inférieures 0, 100, 200, 300 et 400.

Les **limites de classe supérieures** sont les plus grands nombres qui peuvent appartenir aux différentes classes. Le tableau 2-2 a comme limites de classe supérieures 99, 199, 299, 399 et 499.

Les **frontières de classe** sont les nombres utilisés pour séparer les classes. Voici comment on les obtient : trouvez la différence entre la borne supérieure d'une classe et la borne inférieure de la classe suivante. Additionnez la moitié de cette différence à la borne supérieure de classe pour trouver la frontière supérieure de classe et retranchez la moitié de cette différence à la borne inférieure de classe pour trouver la frontière inférieure de classe. Dans le tableau 2-2, les différences sont de une unité, donc on ajoute et on soustrait 0,5 aux bornes pour trouver les frontières. La première classe a –0,5 et 99,5 comme frontières, la seconde classe 99,5 et 199,5 et ainsi de suite.

Les **centres de classe** sont les points situés au milieu de la classe. Chaque centre de classe peut être trouvé en faisant la demi-somme des bornes inférieures et supérieures de la classe. Dans le tableau 2-2, les centres de classe sont 49,5 149,5 249,5 349,5 et 449,5.

La **largeur de classe** est la différence entre deux bornes inférieures consécutives ou deux frontières consécutives. Le tableau 2-2 utilise une largeur de classe de 100.

Les définitions de largeur de classe et de frontières de classe peuvent prêter à confusion. Faites attention à éviter l'erreur classique qui consiste à prendre comme largeur de classe la différence entre la borne supérieure et la borne inférieure. Regardez le tableau 2-2 et notez que la largeur de

évident que la distribution de fréquence des fumeurs est très différente de celles des deux autres groupes. Parce que les deux groupes de non-fumeurs (exposés et non exposés) ont de très hautes fréquences pour la première classe, il est intéressant de comparer plus avant ces jeux de données avec une étude plus fine de leurs valeurs.

Tableau 2-7 Niveaux de cotinine pour les trois groupes

Niveau de cotinine (ng/ml)	Fumeurs	Non fumeurs exposés	Non fumeurs non exposés
0-99	28 %	85 %	95 %
100-199	30 %	5 %	0 %
200-299	35 %	3 %	3 %
300-399	3 %	3 %	3 %
400-499	5 %	0 %	0 %
500-599	0 %	5 %	0 %

Intervalles semi-ouverts Les distributions de fréquences présentées dans cette section sont très « propres » dans ce sens qu'elles considèrent toutes une même largeur de classe. Il est souvent nécessaire d'utiliser des intervalles semi-ouverts comme la catégorie d'âge « 65 ans ou plus ». Il est souvent préférable d'utiliser un tel intervalle qui capture une faible proportion des données de l'échantillon plutôt que d'utiliser de nombreuses classes (comme 65-74, 75-84, 85-94, 95-104) qui contiennent chacune une proportion vraiment très faible des données. Cependant un intervalle semi-ouvert introduit une approximation qui peut devenir gênante quand on doit faire des calculs ou un graphique comme ceux présentés dans la section suivante.

2.2 Exercices

Dans les exercices 1 et 2, identifiez la largeur de classe, les centres de classe et les frontières de classe pour les distributions de fréquences données basées sur le jeu de données 1 de l'annexe B.

1. Pression systolique pour les hommes	Fréquence
90–99	1
100–109	4
110–119	17
120–129	12
130–139	5
140–149	0
150–159	1

2. Cholestérol des hommes	Fréquence
0–199	13
200–399	11
400–599	5
600–799	8
800–999	2
1000–1199	0
1200–1399	1

Tracés tige et feuilles

Un tracé en **tige et feuilles** (*stem-and-leaf*) représente les données en séparant chaque valeur en deux parties : la tige (le chiffre le plus à gauche) et la feuille (le chiffre le plus à droite). L'illustration ci-dessous montre un tracé tige et feuilles pour les mêmes hauteurs (en mètres) de peupliers que celles représentées dans le dotplot du haut de la figure 2.4. Il est facile de voir comment la première hauteur de 3,2 m est séparée en sa tige 3 et sa feuille 2. Chacune des valeurs restantes est découpée de façon similaire. Les feuilles sont ensuite rangées en ordre croissant.

Tracé tige et feuilles

tige (unités)	feuilles (dixièmes)	
3,	29	← les valeurs sont 3,2 et 3,9
4,	4	
5,	4	
6,	334788999	
7,	133367	
8,	0	← la valeur est 8,0

En tournant la page sur la gauche, on peut voir une distribution de ces données. Un grand avantage du tracé tige et feuilles est qu'on peut voir la distribution des données et cependant garder l'information de la liste originale. Si nécessaire, on pourrait reconstruire la liste originale des données. Un autre avantage de cette construction est que cela est un bon et rapide moyen de *trier* les données, le tri des données étant parfois obligatoire pour certaines procédures statistiques (comme pour trouver la médiane, les percentiles, les rangs).

Les lignes de chiffres dans un tracé tige et feuilles sont similaires par nature aux barres d'un histogramme. Une des recommandations pour construire des histogrammes est que le nombre de classes devrait être compris entre 5 et 20 et la même recommandation s'applique aux tracés tige et feuilles pour les mêmes raisons. De meilleurs tracés tige et feuilles sont obtenus en arrondissant d'abord les données originales. On peut aussi *étendre* les tracés pour inclure plus de lignes ou les *condenser* pour avoir moins de lignes en combinant les tiges.

Diagramme de Pareto

Une de ces dernières années, il y a eu 13 800 morts accidentelles parmi les résidents américains âgés de 15 à 24 ans (d'après des données du Conseil national américain de santé). En voici le décompte par catégorie : armes à feu (150), poison (870), véhicules à moteur (10 500), feux et incendies (240), noyades (700), chutes (210) et autres causes (1 130). Bien que la phrase précédente décrive correctement les données, une meilleure compréhension peut être obtenue à l'aide d'un graphique. Un graphique adapté à ces données est le **diagramme de Pareto** qui est un graphique en barres pour des données qualitatives, avec les barres rangées dans l'ordre de leurs fréquences. Comme avec les histogrammes, l'axe vertical dans les diagrammes de Pareto représente les fréquences ou les fréquences relatives. La plus haute barre est sur la gauche et les plus petites sont sur la droite. En arrangeant les barres par ordre de fréquence, le diagramme de Pareto attire l'attention sur les catégories les plus importantes. La figure 2.5 est un diagramme de Pareto qui montre clairement que la catégorie des accidents dus à des véhicules à moteur est de loin la plus importante.

Explorer les données : cherchez des aspects remarquables du graphique qui révèlent des caractéristiques utiles et/ou intéressantes du jeu de données. Dans la figure 2.9 par exemple, on voit que les soldats mouraient plus de mauvais soins à l'hôpital que de blessures.

Comparer les données : construisez des graphiques similaires qui facilitent la comparaison. Par exemple, regardez les dotplots dans cette section et vous verrez que les peupliers traités avec fertilisant et irrigation ont tendance à être plus grands que ceux traités par irrigation seulement.

Dans les exercices 1 et 2, répondez aux questions qui font référence à l'histogramme ci-dessous (figure 2.10) produit par SPSS et qui représente les longueurs (mm) d'œufs de coucous trouvés dans les nids d'autres oiseaux (d'après des données de O. M. Latter et de la bibliothèque de données et histoires DASL).

1. ***Tendance centrale*** Quelle est approximativement la valeur du centre ? C'est-à-dire, quelle longueur d'œuf semble être proche du centre de toutes les longueurs montrées dans le graphique ?

2. ***Pourcentage*** Quel pourcentage des 120 œufs ont une longueur de moins de 21,125 mm ?

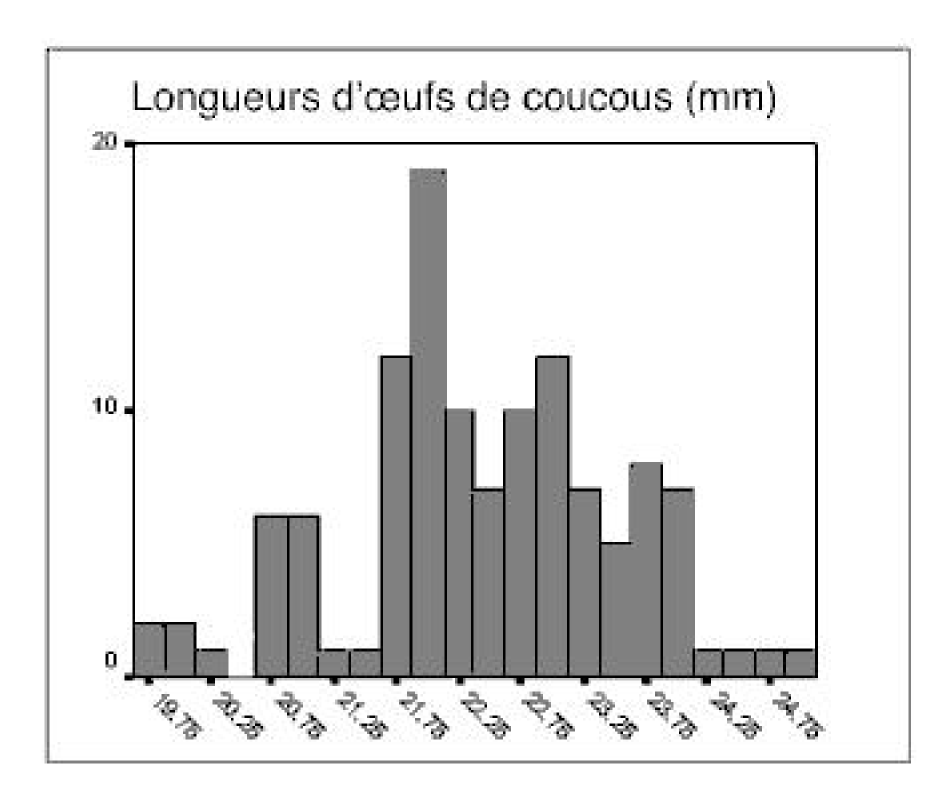


Figure 2.10

Pour l'exercice 3, il faut se reporter au graphique circulaire ci-joint (figure 2.11) des groupes sanguins pour un grand échantillon de personnes (d'après des données du grand programme sanguin de New York).

3. ***Interpréter un graphique circulaire*** Quel est approximativement le pourcentage de personnes du groupe A ? En supposant que le graphique correspond à un échantillon de 500 personnes, combien approximativement de personnes sur ces 500 sont du groupe A ?

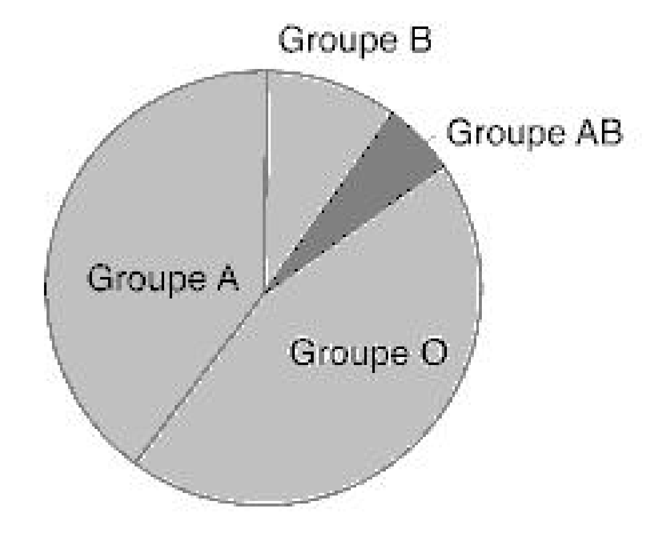
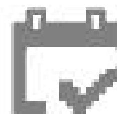


Figure 2.11

Pour trouver la médiane, il faut d'abord trier les données puis utiliser l'une des deux procédures suivantes :

1. si le nombre de valeurs est impair, la médiane est la valeur située exactement au milieu de la liste ;
2. si le nombre de valeurs est pair, la médiane est obtenue en prenant la moyenne des deux valeurs du milieu.



Exemple **Mesure du taux de plomb dans l'air.** On liste ci-dessous des valeurs mesurées de plomb (en μg/m^3). Trouvez la médiane pour cet échantillon.

5,40	1,10	0,42	0,73	0,48	1,10

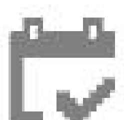
Solution Triez d'abord les valeurs par ordre croissant :

0,42	0,48	0,73	1,10	1,10	5,40

Comme il y a un nombre pair (6) de valeurs, la médiane est obtenue en calculant la moyenne des deux valeurs du milieu, soit 0,73 et 1,10.

$$\text{Médiane} = \frac{0{,}73 + 1{,}10}{2} = \frac{1{,}83}{2} = 0{,}915.$$

Il faut noter que la médiane 0,915 μg/m^3 est très différente de la moyenne 1,538 μg/m^3 qu'on avait trouvée dans l'exemple précédent. Cette grande différence est due à l'effet de la valeur 5,40 sur la moyenne. Si cette valeur extrême était ramenée à 1,20, la moyenne descendrait à 0,838 μg/m^3 alors que la médiane ne changerait pas.



Exemple **Mesure du taux de plomb dans l'air.** Reprenez l'exemple précédent en ajoutant la valeur 0,66 μg/m^3 enregistrée un autre jour.

Solution Triez d'abord les valeurs par ordre croissant :

0,42	0,48	0,73	1,10	1,10	5,40

Comme il y a un nombre impair (7) de valeurs, la médiane est la valeur située exactement au milieu, soit 0,73 μg/m^3.

Après avoir étudié les deux exemples précédents, la procédure pour trouver la médiane devrait être claire. Il devrait aussi être clair que la moyenne est très affectée par les valeurs extrêmes alors que la médiane est peu affectée. Parce que la médiane n'est pas sensible aux valeurs extrêmes, elle est souvent utilisée pour des jeux de données avec un nombre relativement faible de valeurs extrêmes. Par exemple, le bureau US du recensement a rapporté que le revenu médian annuel par foyer était de 36 078 $. La médiane a été utilisée parce qu'il y avait un petit nombre de foyers avec des revenus vraiment très importants.

Tableau 2-9 (suite) Comparaison de la moyenne, de la médiane, du mode et du midrange

Mesure de tendance centrale	Existence	Prend toutes les valeurs en compte ?	Affectée par les valeurs extrêmes	Avantages et inconvénients
Moyenne	existe toujours	oui	oui	utilisée dans tout le livre ; fonctionne bien avec beaucoup de méthodes statistiques
Médiane	existe toujours	non	non	souvent un bon choix s'il y a quelques valeurs extrêmes
Mode	peut ne pas exister ; il peut y avoir plusieurs modes	non	non	adaptée au niveau nominal
Midrange	existe toujours	non	oui	très sensible aux valeurs extrêmes

Commentaires généraux :
– Pour un ensemble de données à peu près symétrique avec un seul mode, la moyenne, la médiane et le midrange ont tendance à être les mêmes.
– Pour un ensemble de données clairement asymétrique, il serait bon de donner à la fois la moyenne et la médiane.
– La moyenne est relativement *fiable*. C'est-à-dire que quand des échantillons sont tirés de la même population, les moyennes des échantillons ont tendance à être plus « consistantes » que les autres mesures de tendance centrale (consistantes au sens où les moyennes des échantillons d'une même population varient moins que les autres mesures de tendance centrale).

Asymétrie

Une comparaison de la moyenne, de la médiane et du mode peut révéler des informations sur la caractéristique d'asymétrie, définie ci-dessous et illustrée par la figure 2.12.

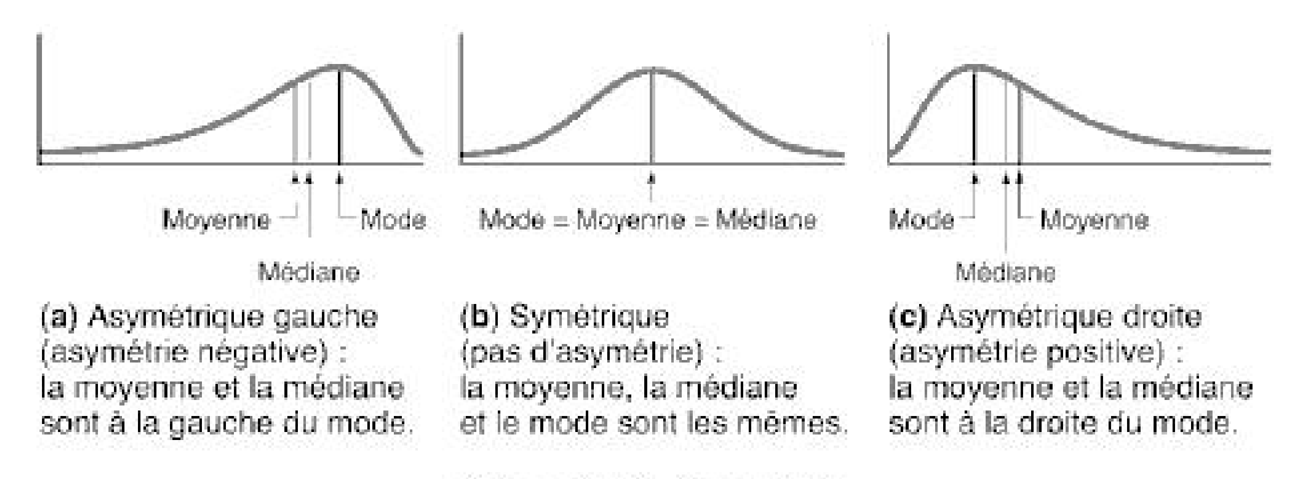
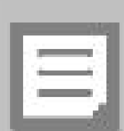


Figure 2.12 Asymétrie



Une distribution de données est **symétrique** si la moitié gauche de l'histogramme est à peu près l'image en miroir de la moitié droite. Une distribution de données est **asymétrique** si elle n'est pas symétrique et si elle s'étend plus sur un côté qu'un autre.

Pour calculer l'étendue, soustrayez simplement la plus petite valeur de la plus grande valeur. Pour les temps d'attente de la file unique, l'étendue est 7 – 4 = 3 min. Les temps d'attente des files multiples ont une étendue de 13 minutes et cette valeur plus grande suggère une plus grande variation.

L'étendue est très facile à calculer mais, parce qu'elle dépend uniquement du minimum et du maximum, elle n'est pas aussi utile que les autres mesures de dispersion qui utilisent toutes les valeurs.

Écart type d'un échantillon

L'écart type est la mesure de variation qui est généralement la plus importante et la plus utile. Nous la définissons maintenant, mais pour bien la comprendre vous devrez étudier la sous-section « Interpréter et comprendre l'écart type » que vous trouverez un peu plus loin dans cette section.



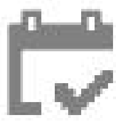
L'**écart type** de l'ensemble des valeurs d'un échantillon est une mesure de dispersion des valeurs autour de la moyenne. Il représente à peu près la déviation moyenne des valeurs par rapport à la moyenne et qui se calcule à l'aide des formules 2.3 ou 2.4.

Formule 2.3 $$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$ Écart type de l'échantillon

Formule 2.4 $$s = \sqrt{\frac{n\sum(x)^2 - (\sum x)^2}{n(n-1)}}$$ Formule avec raccourci pour l'écart type de l'échantillon

Un peu plus loin dans cette section nous discuterons le bien-fondé de ces formules, mais pour l'instant nous vous recommandons d'utiliser la formule 2.3 sur quelques exemples, puis d'apprendre à calculer les écarts types à l'aide d'une calculatrice et avec un programme d'ordinateur. La plupart des calculatrices scientifiques sont conçues pour que vous puissiez saisir une liste de valeurs et obtenir automatiquement l'écart type. Pour l'instant, nous donnons des propriétés importantes qui sont des conséquences de la façon dont l'écart type est défini.

- L'écart type est une mesure de dispersion de toutes les valeurs autour de la moyenne.
- La valeur de l'écart type s est en général positive. Elle est nulle uniquement si toutes les données ont la même valeur. De plus, de plus grandes valeurs de s indiquent une plus grande variation.
- La valeur de l'écart type s peut augmenter de façon importante si on inclut une ou plusieurs valeurs extrêmes (valeurs des données qui sont vraiment très loin des autres).
- Les unités de l'écart type s (comme les minutes, les mètres, les kg et ainsi de suite) sont les mêmes que les unités des données originales.



Exemple Utilisation de la formule 2.4.

Adaptation Supposez que vous ayez les 3 valeurs 1, 3 et 14. Quel en est l'écart type à l'aide de la formule 2.4 ?

la distribution mais nous préférons sacrifier la précision au bénéfice de la simplicité. De plus, on pourrait utiliser trois ou même quatre écarts types au lieu de deux, ce qui était un choix arbitraire. Mais on veut une règle simple qui nous permette d'interpréter les valeurs d'écart type ; d'autres méthodes que nous verrons plus tard permettront de produire des résultats plus précis.

Recette de l'étendue

Pour estimer une valeur de l'écart type s : pour une approximation rapide de l'écart type, utilisez

$$s \approx \frac{\text{étendue}}{4}$$

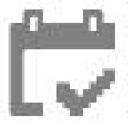
où étendue = maximum – minimum.

Pour interpréter une valeur connue de l'écart type s : si l'écart type s est connu, utilisez, pour trouver des approximations rapides du minimum « usuel » et du maximum « usuel » des valeurs de l'échantillon, les expressions :

minimum « usuel » = moyenne – 2 × écart type ;

maximum « usuel » = moyenne + 2 × écart type.

Quand vous calculez l'écart type à l'aide de la formule 2.4, vous pouvez utiliser la recette de l'étendue pour vérifier votre résultat mais vous devez réaliser que même si l'approximation peut vous amener au voisinage de la réponse, elle peut aussi en être assez éloignée.



Exemple Niveaux de cotinine des fumeurs. Utilisez la recette de l'étendue pour trouver une approximation rapide de l'écart type de l'échantillon des niveaux de cotinine pour les 40 fumeurs du tableau 2-1.

Solution À l'aide de cette recette on calcule l'étendue et on divise par 4. Si on parcourt les données, on voit que le minimum est 0 et le maximum 491, donc l'étendue est 491. L'écart type est approximé ainsi :

$$s \approx \frac{\text{étendue}}{4} = \frac{491}{4} = 122{,}75 \approx 123.$$

Interprétation Ce résultat est très proche de la valeur exacte, qui est 119,5, obtenue avec la formule 2.3 ou 2.4. Il ne faut pas espérer que la recette marchera aussi bien dans tous les autres cas.

L'exemple suivant est particulièrement important pour illustrer un moyen d'*interpréter* la valeur de l'écart type.

Exemple Circonférence crânienne des filles. D'anciens résultats d'une enquête nationale américaine de santé suggèrent que la circonférence crânienne d'enfants (filles) de 2 mois est en moyenne de 40,05 cm avec un écart type de 1,64 cm. Utilisez la recette d'étendue pour en trouver le minimum usuel et le maximum usuel. Ces résultats pourraient être utilisés par un médecin pour détecter une circonférence crânienne « inhabituelle » et qui pourrait être due à une pathologie comme l'hydrocéphalie. Déterminez ensuite si 42,6 cm pourrait être considérée comme « inhabituelle ».

Pourquoi diviser par $n - 1$?

Après avoir trouvé les valeurs individuelles $(x - \overline{x})^2$, nous les combinons en prenant leur somme puis nous en calculons la moyenne en divisant par $n - 1$. On divise par $n - 1$ parce qu'il y a seulement $n - 1$ valeurs indépendantes. C'est-à-dire que pour une moyenne donnée, on peut affecter des valeurs arbitraires à seulement $n - 1$ valeurs avant que la dernière valeur ne soit déterminée. On peut montrer que la division par $n - 1$ rend les variances d'échantillons s^2 plus proches de la variance de la population σ^2 alors que la division par n a tendance à générer des variances d'échantillons qui sous-estiment la variance de la population σ^2.

Une conséquence importante du fait que l'écart type utilise la racine carrée de la somme de carrés est que l'écart type est exprimé dans les mêmes unités de mesure que les données originales. Par exemple, si les temps d'attente sont en minutes, l'écart type de ces temps sera aussi en minutes.

Après avoir étudié cette section, vous devriez avoir compris que l'écart type est une mesure de dispersion entre les valeurs. À partir de données d'échantillon vous devriez être capable de calculer la valeur de l'écart type et d'interpréter les valeurs d'écart type que vous calculez. Vous devriez savoir que pour des jeux de données courants, il est inhabituel pour une valeur d'être à plus de 2 ou 3 écarts types de la moyenne.

2.5 Exercices

Dans les exercices 1 à 4, trouvez l'étendue, la variance et l'écart type pour les données d'échantillon fournies. Il s'agit des mêmes données que pour la section II.4 où on cherchait des mesures de tendance centrale. Ici, on cherche des mesures de dispersion.

1. ***Utilisation du tabac dans les films pour enfants*** Dans « Tobacco and Alcohol Use in G-Rated Children's Animated Films » par Goldstein, Sobel et Newman (*Journal of the American Medical Association*, vol. 281, n°12), la durée (en secondes) des scènes montrant l'utilisation du tabac a été enregistrée pour les films d'animation des studios Universal. Six de ces durées sont indiquées ci-dessous.

0 223 0 176 0 548

2. ***Indice de masse corporelle*** Durant l'enquête nationale américaine de santé, l'indice de masse corporelle (IMC) a été mesuré pour un échantillon de femmes. Quelques-unes des valeurs du jeu de données 1 de l'annexe B sont listées ci-dessous. À partir de ces données d'échantillon, est-ce qu'un IMC de 34,0 peut être considéré comme « inhabituel » ? Et pourquoi ? *Aide* : utilisez la recette de l'étendue.

19,6	23,8	19,6	29,1	25,2	21,4	22,0	27,5
33,5	20,6	29,9	17,7	24,0	28,9	37,7	

3. ***Accidents mortels chez les motards*** On liste ci-dessous les âges de motards quand ils ont été mortellement blessés dans des accidents de la circulation (à partir de données du ministère américain des Transports). Quelle comparaison peut-on faire entre la dispersion de ces âges avec celle des âges des motards dans la population générale ?

17	38	27	14	18	34	16	42	28
24	40	20	23	31	37	21	30	25

4. ***Mesures de tension*** Quatorze étudiants en deuxième année de médecine à l'hôpital de Bellevue ont mesuré la tension d'une même personne. Les valeurs systoliques (en mmHg) sont listées ci-dessous. Que suggèrent les mesures de dispersion de ces données sur leur précision ?

138	130	135	140	120	125	120
130	130	144	143	140	130	150

Quartiles et percentiles

Depuis la section II.4, nous savons que la médiane d'un ensemble de données est la valeur du milieu et donc que 50 % des valeurs sont inférieures ou égales à la médiane et que 50 % lui sont supérieures ou égales. De la même façon que la médiane divise les données en deux parties égales, les trois quartiles, notés Q_1, Q_2 et Q_3, partagent les données triées en quatre parties égales.

Voici les descriptions des trois quartiles :

Q_1 **(premier quartile)** : sépare les premiers 25 % des données triées des autres 75 %. Pour être plus précis, au moins 25 % des données triées sont inférieures ou égales à Q_1 et au moins 75 % des valeurs sont supérieures ou égales à Q_1.

Q_2 **(deuxième quartile)** : c'est la même chose que la médiane, sépare les premiers 50 % des données triées des autres 50 %.

Q_3 **(troisième quartile)** : sépare les premiers 75 % des données triées des autres 25 %. Pour être plus précis, au moins 75 % des données triées sont inférieures ou égales à Q_3 et au moins 25 % des valeurs sont supérieures ou égales à Q_3.

Nous décrirons une procédure pour trouver les quartiles après avoir discuté des percentiles. Il n'y a pas d'accord général sur une seule et même procédure pour calculer les quartiles et des programmes informatiques différents donnent souvent des résultats différents. Par exemple, si vous utilisez les données 1, 3, 6, 10, 15, 21, 28 et 36, vous trouverez ces résultats :

	Q_1	Q_2	Q_3
STATDISK	4,5	12,5	24,5
SPSS	3,75	12,5	26,25
SAS	4,5	12,5	24,5
Excel	5,25	12,5	22,75

Si vous utilisez une calculatrice ou un ordinateur pour les exercices impliquant des quartiles, vous risquez de trouver des résultats légèrement différents des réponses données à la fin du livre.

De la même façon qu'il y a trois quartiles séparant les données en quatre parties égales, il y a aussi 99 **percentiles**, notés P_1, P_2… P_{99}, qui partitionnent les données en 100 groupes avec à peu près 1 % des données dans chaque groupe. Les quartiles et les percentiles sont des exemples de quantiles – ou fractiles – qui partitionnent les données en groupes avec approximativement autant de valeurs.

La procédure pour trouver le percentile qui correspond à une valeur particulière x est assez simple, comme indiqué dans l'expression suivante.

$$\text{Percentile de valeur } x = \frac{\text{Nombre de valeurs inférieures à } x}{\text{Nombre total de valeurs}} \times 100.$$

Dans les exercices 6 et 7, utilisez les 40 niveaux triés de cotinine des fumeurs listés dans le tableau 2-10. Trouvez le percentile correspondant au niveau de cotinine indiqué.

6. 149

7. 35

Dans les exercices 8 à 11, utilisez les 40 niveaux triés de cotinine des fumeurs listés dans le tableau 2-10. Trouvez le percentile ou le quartile indiqué.

8. P_{20}

9. P_{75}

10. P_{33}

11. P_1

12. ***Niveaux de cotinine des fumeurs*** Utilisez les 40 niveaux triés de cotinine des fumeurs listés dans le tableau 2-10.

a. Trouvez la distance interquartile.

b. Trouvez le midquartile.

c. Trouvez l'étendue 10 %-90 %.

d. Est-ce que $P_{50} = Q_2$? Si oui, est-ce que P_{50} vaut toujours Q_2 ?

e. Est-ce que $Q_2 = (Q_1 + Q_3)/2$? Si oui, est-ce que Q_2 vaut toujours $(Q_1 + Q_3)/2$?

II.7 Analyse des données exploratoires

Ce chapitre présente les outils de base pour décrire, explorer et comparer les données, et le point central de cette section est l'exploration des données. Nous commençons cette section par la définition de l'analyse des données exploratoires, puis nous introduirons les valeurs extrêmes, les résumés en 5 nombres et les boîtes à moustaches.



L'analyse des données exploratoires est le procédé par lequel on utilise des outils statistiques (comme les graphiques, les mesures de tendance centrale et de dispersion) pour explorer les jeux de données de façon à comprendre leurs caractéristiques importantes.

Rappelez-vous que dans la section II.1 nous avions listé cinq caractéristiques importantes des données en commençant par (1) la *tendance centrale*, (2) la *dispersion* et (3) la nature de la *distribution*. On peut appréhender ces caractéristiques en calculant les valeurs de la moyenne et de l'écart type et en construisant un histogramme. Il est en général important d'aller plus loin pour identifier des traits remarquables, et spécialement ceux qui pourraient perturber fortement les résultats et les conclusions. Un de ces traits est la présence de valeurs extrêmes.

Valeurs extrêmes

Une **valeur extrême** est une valeur située très loin de toutes les autres valeurs. Par rapport aux autres valeurs, cette donnée est *très à l'écart*. Quand on explore un jeu de données, on doit considérer les valeurs extrêmes parce qu'elles peuvent révéler des informations importantes et affecter fortement les valeurs de la moyenne et de l'écart type, de même qu'elles peuvent déformer un histogramme.

La description précédente des boîtes à moustaches présentait les boîtes à moustaches **squelettales** (ou **régulières**), mais quelques logiciels statistiques fournissent des **boîtes à moustaches modifiées** qui montrent les valeurs extrêmes comme des points spéciaux (comme dans le graphique de SPSS pour les niveaux de cholestérol des femmes). Par exemple, Minitab utilise des astérisques pour identifier les points qui sont exceptionnels parce qu'ils sont plus grands ou plus petits que la plupart des autres points. Le critère spécifique de Minitab est de mettre des astérisques pour les points qui représentent des valeurs plus petites que $Q_1 - 1{,}5 \times (Q_3 - Q_1)$ ou plus grandes que $Q_1 + 1{,}5 \times (Q_3 - Q_1)$ comme dans l'exemple suivant. Une autre approche est d'utiliser des petits cercles pleins pour les « *faibles* » *valeurs extrêmes* et des petits cercles vides pour les « *fortes* » *valeurs extrêmes* définies comme suit.

Faibles valeurs extrêmes (tracées comme des *petits cercles pleins*) : valeurs en dessous de Q_1 ou au-dessus de Q_3 de plus de $1{,}5 \times (Q_3 - Q_1)$ mais de moins de $3 \times (Q_3 - Q_1)$.

Fortes valeurs extrêmes (tracées comme des *petits cercles vides*) : valeurs en dessous de Q_1 ou au-dessus de Q_3 de plus de $3 \times (Q_3 - Q_1)$.

EXEMPLE Est-ce que les hommes et les femmes ont les mêmes rythmes cardiaques ? Il a souvent été mentionné qu'il y a des différences physiologiques entre les hommes et les femmes. Les hommes ont tendance à être plus grands et plus lourds que les femmes. Mais y a-t-il une différence pour le rythme cardiaque ? Le jeu de données 1 de l'annexe B liste les rythmes cardiaques pour un échantillon de 40 hommes et un échantillon de 40 femmes. Plus tard dans ce livre nous décrirons des méthodes statistiques importantes qui seront utilisées pour tester formellement des différences, mais pour l'instant nous allons explorer les données pour voir ce que nous pouvons apprendre. Même si nous savions comment utiliser ces méthodes, il serait sage d'explorer les données avant d'utiliser les procédures formelles.

SOLUTION Commençons par examiner les éléments clés de tendance centrale, dispersion, valeurs extrêmes et changement (ceux de la liste « TDDVT » introduite dans la section II.1). Les affichages suivants (figures 2.22, 2.23 et 2.24) montrent des graphiques créés par Minitab, SPSS et SAS.

On liste ci-dessous les mesures de tendance centrale (moyenne), de dispersion (écart type) et le résumé en 5 nombres pour les rythmes cardiaques du jeu de données 1.

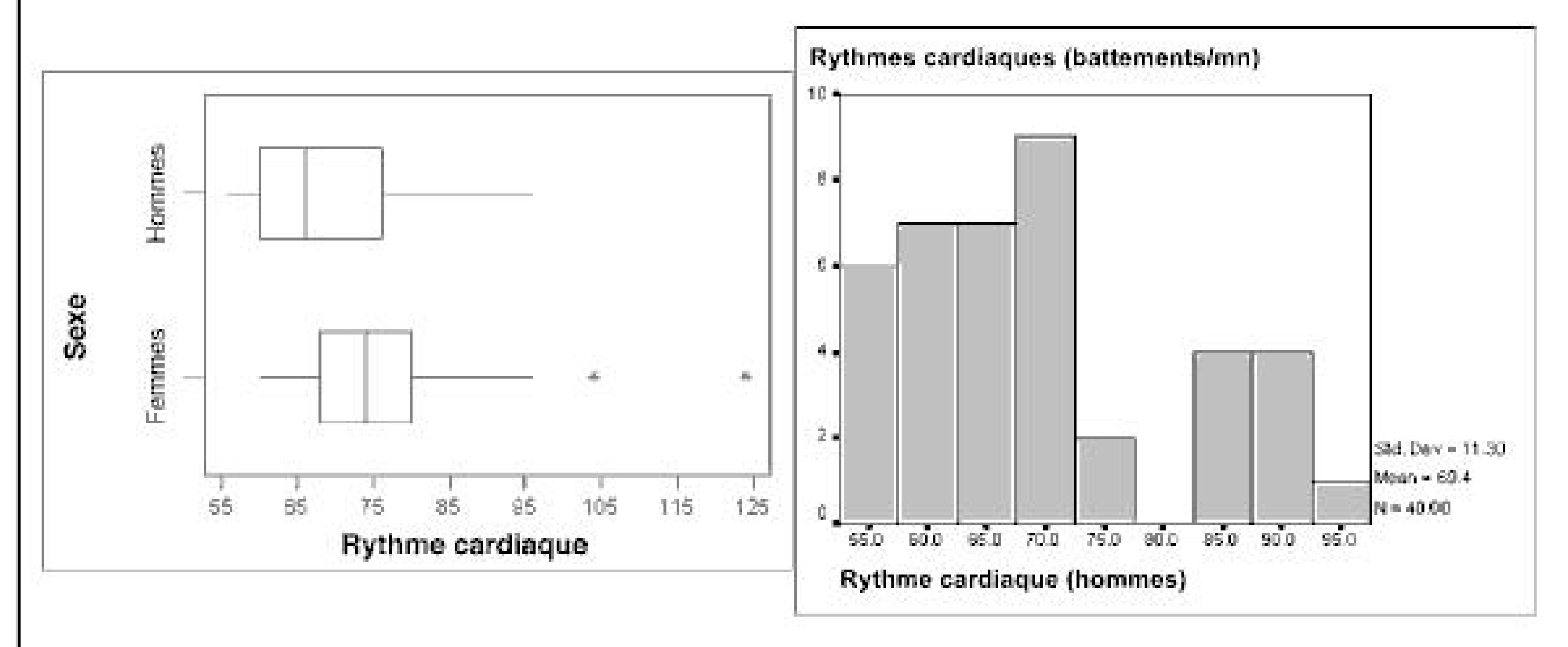


Figure 2.22 MINITAB **Figure 2.23 SPSS**

3 Estimations et tailles d'échantillons avec un échantillon

Problème du chapitre 3

Est-ce que Mendel avait tort ?

Quand Gregor Mendel a mené ses fameuses expériences de génétiques avec des pois, un des échantillons des croisements a été obtenu en croisant des pois à gousses vertes et des pois à gousses jaunes. Cette lignée comportait 580 pois. Parmi ces pois, 428 avait des gousses vertes et 152 des gousses jaunes. À partir de sa théorie des gènes, Mendel s'attendait à ce que 25 % des pois aient des gousses jaunes. Toutefois, avec 428 gousses vertes et 152 gousses jaunes, le pourcentage de gousses jaunes est de 26,2 %. Comment peut-on expliquer cette différence ? Cette différence est-elle suffisamment importante pour suggérer que les 25 % de Mendel sont incorrects ? Si nous ignorions la théorie de Mendel et que nous utilisions seulement les résultats de l'expérience, quelle estimation de gousses jaunes pourrions-nous attendre d'expériences similaires ? Et que pouvons-nous dire sur la précision de cette estimation ?

Ce chapitre présente les concepts statistiques nécessaires pour répondre à de telles questions. Nous analyserons les résultats de l'expérience de Mendel et nous apprendrons beaucoup à propos de l'estimation des paramètres d'une population en général. Bien que ces données impliquent l'estimation de la proportion d'une population, ce chapitre considérera aussi l'estimation de la moyenne et de la variance d'une population.

Valeurs critiques

Les méthodes de ce chapitre et de nombreuses autres méthodes statistiques des chapitres suivants incluent l'utilisation d'un score normalisé, noté z, qui permet de distinguer les statistiques d'échantillon qui peuvent vraisemblablement survenir de celles qui ne le peuvent pas.

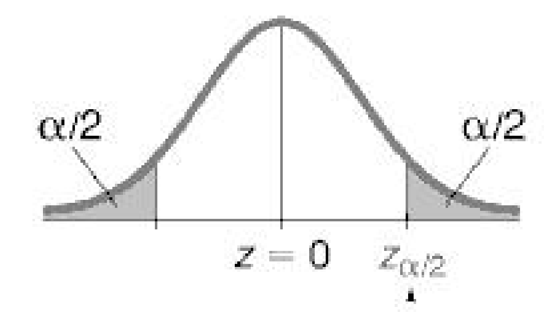


Figure 3.1 Valeur critique $z_{\alpha/2}$ pour la loi normale standard

Un tel score-z est appelé une *valeur critique* (définition ci-dessous). Les valeurs critiques sont basées sur les observations suivantes :

1. sous certaines conditions, la distribution de l'échantillon peut être approximée par la loi normale comme dans la figure 3.1 ;
2. les proportions d'échantillon ont une chance relativement faible (avec une probabilité notée α) de tomber dans l'une des deux queues gris foncé de la figure 3.1 ;
3. si on note $\alpha/2$ l'aire de chaque queue gris foncé, on voit qu'il y a une probabilité totale α que la proportion de l'échantillon tombe dans l'une de ces deux zones gris foncé ;
4. il y a donc une probabilité de $1 - \alpha$ que la proportion de l'échantillon tombe dans la zone intérieure gris clair de la figure 3.1 ;
5. le score-z qui sépare la région de la queue droite est noté couramment $z_{\alpha/2}$ et on s'y réfère comme la *valeur critique* parce que c'est la frontière qui sépare les proportions d'échantillon qui peuvent vraisemblablement survenir de celles qui ne le peuvent pas.

Ces observations peuvent être formalisées comme suit.

Notations pour la valeur critique

La valeur critique $z_{\alpha/2}$ est la valeur z positive à la frontière verticale qui sépare une aire de $\alpha/2$ dans la queue droite de la loi normale standard. La valeur $-z_{\alpha/2}$ est la valeur à la frontière verticale qui sépare une aire de $\alpha/2$ dans la queue gauche. L'indice $\alpha/2$ est juste là pour rappeler que le score-z délimite des aires de $\alpha/2$.

Une **valeur critique** est un nombre sur la frontière séparant les statistiques d'échantillon qui peuvent vraisemblablement survenir de celles qui ne le peuvent pas. Le nombre $z_{\alpha/2}$ est une valeur critique qui est un score-z et délimite des aires de $\alpha/2$ pour la loi normale standard (voir figure 3.1).

appartient à l'intervalle [0,226 ; 0,298]. Cela est illustré par la figure 3.3. Le premier intervalle de confiance de cette figure est celui avec l'expérience du problème introductif mais les 19 autres représentent des échantillons hypothétiques. Avec un niveau de confiance de 95 %, on s'attend à ce que 19 des 20 échantillons contiennent la vraie valeur de p et la figure 3.3 le montre, avec 19 intervalles qui contiennent p et un seul qui ne la contient pas.

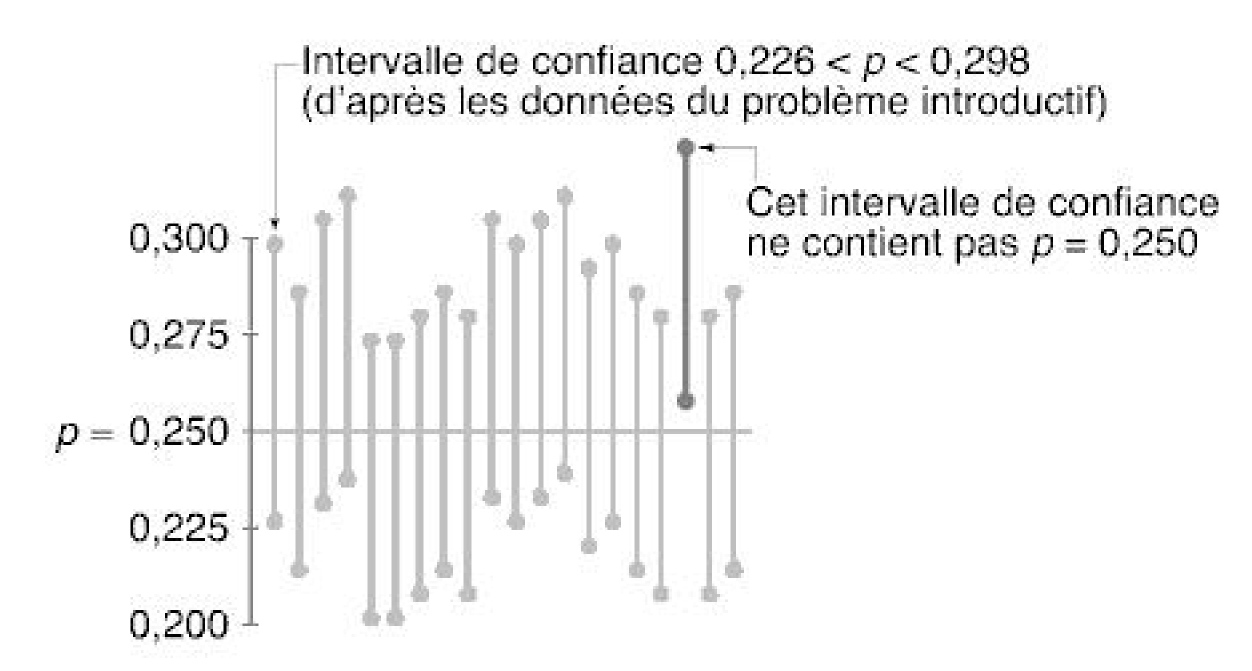


Figure 3.3 Intervalles de confiance pour 20 échantillons différents

Raison d'être de la marge d'erreur. Parce que la loi de l'échantillon est approximativement normale (car les conditions $np \geqslant 5$ et $nq \geqslant 5$ sont satisfaites), on peut utiliser des théorèmes probabilistes pour conclure que μ et σ sont donnés par $\mu = np$ et $\sigma = \sqrt{npq}$. Ces deux paramètres sont en rapport avec n essais, mais nous les convertissons en « par essai » en divisant par n comme suit :

Moyenne des proportions d'échantillon : $\mu = \dfrac{np}{n} = p.$

Écart type des proportions d'échantillon : $\sigma = \dfrac{\sqrt{npq}}{n} = \sqrt{\dfrac{pq}{n}}.$

Le premier résultat peut sembler trivial parce que nous avions déjà stipulé que la vraie proportion de la population était p. Le second résultat est moins évident et il est utile pour décrire la marge d'erreur E, mais nous remplaçons le produit pq par $\hat{p}\hat{q}$ parce que nous ne connaissons pas encore la valeur de p. La formule 3.1 pour la marge d'erreur reflète le fait que $\hat{p}$ a la probabilité $1 - \alpha$ d'être à moins de $z_{\alpha/2}\sqrt{pq / n}$ de p. L'intervalle de confiance pour p fourni précédemment reflète le fait qu'il y a une probabilité $1 - \alpha$ que $\hat{p}$ diffère de p de moins que la marge d'erreur $E = z_{\alpha/2}\sqrt{pq / n}$.

Déterminer la taille d'échantillon

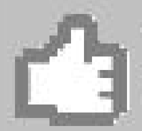
Supposez que l'on veuille collecter des données en vue d'estimer la proportion d'une population. Comment savons-nous combien d'items d'échantillons il faut obtenir ? Par exemple, supposez que nous voulions estimer la proportion de filles nées pour des jumeaux, triplés, quadruplés, quintuplés et sextuplés. *Combien* faut-il observer de tels enfants pour avoir une estimation raisonnable ?

III.3 Estimer la moyenne d'une population : σ connu

Dans la section III.2 nous avons introduit l'estimation ponctuelle et l'intervalle de confiance comme des outils pour trouver la proportion d'une population à l'aide de la proportion de l'échantillon. Nous avons aussi montré comment déterminer la taille minimale requise de l'échantillon pour estimer une proportion de population. Dans cette section, nous allons à nouveau discuter d'estimation ponctuelle, d'intervalle de confiance et de détermination de taille d'échantillon mais notre but maintenant est d'estimer la moyenne μ d'une population. Les estimations de moyennes de population sont souvent extrêmement importantes. Par exemple, d'importantes questions comme celles-ci peuvent être traitées à l'aide des méthodes de cette section et de la suivante :

- quelle est la durée de vie moyenne des aigles chauves aux États-Unis ?
- quel est le poids moyen des éléphants au Kenya ?
- quelle est la production moyenne de lait de vache obtenue dans l'état de New York ?

Les conditions suivantes s'appliquent aux méthodes introduites dans cette section (il y a d'autres conditions pour d'autres procédures similaires).



Conditions requises pour estimer μ quand σ est connu

1. L'échantillon est un échantillon aléatoire simple (tous les échantillons de la même taille ont la même chance d'être sélectionnés).
2. La valeur de l'écart type σ de la population est connue.
3. L'une des deux ou les deux conditions suivantes sont satisfaites : la population est normalement distribuée ou $n > 30$. Parce qu'on ne connaît pas toutes les valeurs de la population, on peut tester la normalité à l'aide d'outils comme les histogrammes, les tracés de quantiles normaux et des valeurs extrêmes trouvées pour les données *d'échantillon*.

Quand on utilise les procédures de cette section pour estimer la moyenne inconnue μ d'une population, les conditions ci-dessous indiquent que nous devons connaître la valeur de l'écart type σ de la population. Il serait cependant très inhabituel de pouvoir connaître σ sans connaître μ. Après tout, la seule façon de connaître σ, c'est de le calculer à partir de toutes les valeurs de la population, donc le calcul de μ serait également possible et si on peut connaître la vraie valeur de μ, il n'y a aucun besoin de l'estimer. Bien que les méthodes pour l'intervalle de confiance de cette section ne soient pas très réalistes, elles révèlent les concepts de base d'importants raisonnements statistiques et elles sont la base pour la détermination de la taille de l'échantillon vue plus tard dans cette section.

Condition de normalité. Dans cette section, nous utilisons la condition qu'il faut disposer d'un échantillon aléatoire simple, que la valeur de σ est connue et que soit la population est normale, soit $n > 30$. Techniquement, la population n'a pas besoin de suivre une loi exactement normale mais elle devrait être approximativement normale, ce qui signifie que sa distribution est en gros symétrique avec un seul mode et pas de valeurs extrêmes. Testez la normalité en construisant l'histogramme des données de l'échantillon et décidez ensuite s'il est à peu près en cloche. Un tracé de quantiles normaux peut aussi être construit mais les méthodes de cette section sont dites **robustes**, ce qui signifie que ces méthodes ne sont pas fortement affectées si on s'éloigne de la normalité, pour peu qu'on ne s'en écarte pas de façon trop extrême. On peut en général considérer que la population est normale lorsque les données de l'échantillon confirment qu'il n'y a pas de valeurs extrêmes et que l'histogramme a une forme qui n'est pas trop éloignée de celle

Solution

On doit d'abord vérifier que les conditions requises sont satisfaites. À partir de l'énoncé, on sait que l'échantillon est un échantillon aléatoire simple. On sait aussi qu'on peut supposer que $\sigma = 0{,}34$ °C. La troisième condition est d'avoir « soit une loi normale, soit $n > 30$ ». Parce que la taille de l'échantillon n est 106, on a $n > 30$ donc on n'a pas besoin de tester si l'échantillon suit une loi normale. Les conditions requises sont satisfaites et on peut donc appliquer la formule 3.4. ❑

a. Le niveau de confiance 95 % correspond à $\alpha = 0{,}05$ donc $z_{\alpha/2} = 1{,}96$. La marge d'erreur E vaut donc (les décimales supplémentaires seront utilisées pour minimiser les erreurs d'arrondi des bornes de l'intervalle de confiance de la partie b.) :

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{0{,}34}{\sqrt{106}} = 0{,}06472649.$$

b. Avec $\bar{x} = 36{,}78$ et $E = 0{,}06472649$, voici l'intervalle de confiance :

$$\bar{x} - E < \mu < \bar{x} + E$$

$$36{,}78 - 0{,}06472649 < \mu < 36{,}78 + 0{,}06472649$$

$$36{,}72 < \mu < 36{,}84 \text{ (arrondi à deux décimales comme pour } \bar{x}\text{)}$$

Interprétation Ce résultat pourrait aussi être présenté comme $36{,}78 \pm 0{,}06$. À partir de l'échantillon avec $n = 106$, $\bar{x} = 36{,}78$ °C et en supposant que $\sigma = 0{,}34$ °C, l'intervalle de confiance à 95 % pour la moyenne μ de la population est $36{,}72 < \mu < 36{,}84$. Cela signifie que si on sélectionnait de nombreux échantillons de taille 106 et qu'on construisait les intervalles de confiance correspondants, 95 % d'entre eux contiendraient effectivement la vraie valeur μ de la moyenne de la population.

Il est à noter que l'intervalle de confiance [36,72 ; 36,84] ne contient pas 37,00 qui est la valeur communément admise pour le corps humain. À partir de ces résultats, il semble peu probable que 37,00 soit la température moyenne correcte du corps humain pour la population totale.

Raison d'être de l'intervalle de confiance. L'idée de base derrière la construction des intervalles de confiance est liée au théorème de la limite centrale qui indique que si on a un échantillon aléatoire simple d'une distribution normale ou un échantillon aléatoire simple de taille $n > 30$ de n'importe quelle population, la distribution des moyennes d'échantillon est approximativement normale, de moyenne μ et d'écart type $\sigma/\sqrt{n}$. Le format de l'intervalle de confiance est issu de l'équation utilisée dans ce théorème. Dans l'expression $z = (\bar{x} - \mu_{\bar{x}})/\sigma_{\bar{x}}$, il faut remplacer $\sigma_{\bar{x}}$ par $\sigma/\sqrt{n}$ et trouver μ, soit :

$$\mu = \bar{x} - z\frac{\sigma}{\sqrt{n}}$$

À l'aide de valeurs positives et négatives de z on trouve les bornes de l'intervalle de confiance utilisé.

Considérons le cas spécifique du niveau de confiance 95 %, soit $\alpha = 0{,}05$ et $z_{\alpha/2} = 1{,}96$. Pour ce cas, il y a une probabilité 0,05 que la moyenne de l'échantillon soit distante de μ de plus de 1,96 écart type (ou $z_{\alpha/2}\sigma/\sqrt{n}$ que l'on note E). Inversement, il y a une probabilité de 0,95 que la moyenne de

Détermination de la taille de l'échantillon. *Dans les exercices 7 et 8, utilisez la marge d'erreur fournie, le niveau de confiance et l'écart type σ de la population pour trouver la taille d'échantillon minimale requise pour estimer la moyenne μ inconnue de la population.*

7. Marge d'erreur : 125 $, niveau de confiance : 95 %, σ = 500 $.

8. Marge d'erreur : 5 min, niveau de confiance : 90 %, σ = 48 min.

Interprétation des résultats. *Dans les exercices 9 et 10, reportez-vous à l'affichage ci-desssous (figure 3.5) d'un intervalle de confiance à 95 % construit à partir des méthodes de cette section. L'affichage des résultats d'échantillon correspond à un échantillon de 80 niveaux de cholestérol sélectionnés aléatoirement pour 80 adultes.*

```
ZInterval
 (262.09,374.11)
 x̄=318.1
 n=80
```

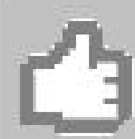
Figure 3.5

9. Identifiez l'estimation ponctuelle de la moyenne μ de la population.

10. Exprimez l'intervalle de confiance sous la forme $\bar{x} \pm E$.

III.4 Estimer la moyenne d'une population : σ inconnu

Dans la section III.3 nous avons présenté des méthodes pour construire un intervalle de confiance qui estime la moyenne μ inconnue d'une population mais on ne considérait que des cas pour lesquels l'écart type σ de la population était connu. Dans cette section, nous allons présenter une méthode pour construire un intervalle de confiance qui estime la moyenne μ inconnue d'une population sans la condition que σ est connu. La procédure usuelle consiste à collecter les données d'échantillon et à trouver les statistiques n, $\bar{x}$ et s. Parce que les méthodes de cette section sont basées sur ces statistiques et que σ n'est pas requis, les méthodes de cette section sont très réalistes, pratiques et souvent utilisées.



Conditions requises pour estimer μ quand σ est inconnu

1. L'échantillon est un échantillon aléatoire simple.
2. Soit la population est normalement distribuée, soit $n > 30$.

Comme dans la section III.3, la condition de normalité n'est pas une condition stricte. On peut considérer qu'une distribution est normale en confirmant qu'il n'y a pas de valeurs extrêmes et que l'histogramme a une forme qui n'est pas trop éloignée de celle d'une loi normale. Comme dans la section III.3, la condition $n > 30$ est couramment utilisée comme repère mais la taille minimale de l'échantillon dépend de la façon dont la distribution de la population s'éloigne de la loi normale. On utilise la condition $n > 30$ comme une justification pour traiter la distribution des moyennes d'échantillon comme une loi normale. La distribution des moyennes d'échantillon $\bar{x}$ suit exactement une loi normale de moyenne μ et d'écart type $\sigma/\sqrt{n}$ quand la population suit une loi normale de moyenne μ et d'écart type σ. Si la population ne suit pas une loi normale, de grands échantillons fournissent des moyennes d'échantillon qui suivent *approximativement* une loi normale de moyenne μ et d'écart type $\sigma/\sqrt{n}$.

Propriétés importantes de la loi *t* de Student

1. La loi t de Student est différente pour différentes tailles d'échantillon (voir figure 3.6 pour les cas $n = 3$ et $n = 12$).
2. La loi t de Student a la même forme en cloche symétrique que la loi normale mais elle reflète une plus grande variabilité (avec des distributions plus larges) qui est attendue pour de petits échantillons.
3. La loi t de Student a une moyenne de $t = 0$ (tout comme la loi normale a une moyenne de $z = 0$).
4. L'écart type de la loi t de Student varie avec la taille de l'échantillon mais il est plus grand que 1 (contrairement à la loi normale pour laquelle $\sigma = 1$).
5. Au fur et à mesure que la taille de l'échantillon n augmente, la loi t de Student se rapproche de la loi normale.

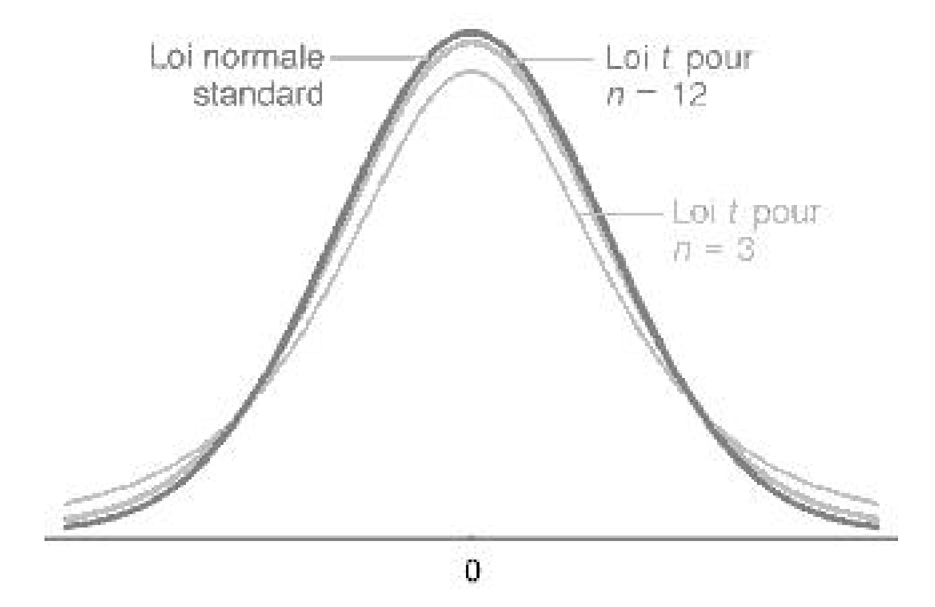


Figure 3.6 Loi *t* de Student pour $n = 3$ et $n = 12$

Ce qui suit est un résumé des conditions d'utilisation de la loi t plutôt que la loi normale. Ces mêmes conditions s'appliqueront au chapitre suivant.

Conditions pour utiliser la loi *t* de Student

1. σ est inconnu.
2. Soit la population suit une loi essentiellement normale, soit $n > 30$.

Choisir la distribution appropriée

Il est parfois difficile de décider s'il faut choisir la loi normale z ou la loi t de Student. L'organigramme de la figure 3.7 et le tableau 3-1 qui l'accompagne résument tous les deux les points clés à considérer quand on construit les intervalles de confiance pour estimer μ, la moyenne de la population. Dans la figure 3.7 ou le tableau 3-1, on notera que si on a un petit ($n \leqslant 30$) échantillon tiré d'une distribution qui diffère beaucoup d'une loi normale, on ne peut pas utiliser les méthodes décrites dans ce chapitre. Une solution est d'utiliser des méthodes non paramétriques (voir le chapitre 9). Une autre solution est d'utiliser un ordinateur avec une méthode de test par permutations (ou *bootstrap*). Dans ces deux approches, aucune hypothèse n'est faite sur la population originale.

Estimation ponctuelle de μ : $$\bar{x} = \frac{(\text{limite supérieure}) + (\text{limite inférieure})}{2}$$

Marge d'erreur : $$E = \frac{(\text{limite supérieure}) - (\text{limite inférieure})}{2}$$

Exemple **Rythme cardiaque des femmes.** Si on analyse avec SPSS le rythme cardiaque des femmes du jeu de données 1 de l'annexe B, on obtient l'affichage qui suit (figure 3.9). Utilisez-le pour trouver l'estimation ponctuelle $\bar{x}$ et la marge d'erreur E. L'échantillon de rythmes cardiaques est tiré aléatoirement d'une grande population de femmes.

Descriptives

			Statistic	Std. Error
Pulse rate (beats/min)	Mean		76.30	1.976
	95% Confidence Interval for Mean	Lower Bound	72.30	
		Upper Bound	80.30	

Figure 3.9

Solution Dans les calculs suivants, les résultats sont arrondis avec une décimale supplémentaire.

$\bar{x}$ = (limite supérieure) + (limite inférieure)/2
= (80,30 + 72,30)/2 = 76,30 pulsations par minute.

E = (limite supérieure) – (limite inférieure)/2
= (80,30 – 72,30)/2 = 4,00 pulsations par minute.

Utiliser les intervalles de confiance pour décrire, explorer ou comparer les données

Dans certains cas, on peut vouloir utiliser un intervalle de confiance pour estimer un paramètre d'une population. Pour les températures corporelles utilisées dans cette section, un but important pourrait être d'estimer la température corporelle moyenne d'adultes en bonne santé et nos résultats suggèrent fortement que la valeur couramment utilisée (37,0 °C) est incorrecte (parce qu'on est sûr à 95 % que les valeurs 36,78 et 36,84 contiennent la vraie moyenne de la population). Dans d'autres cas, un intervalle de confiance peut être un outil parmi d'autres pour décrire, explorer ou comparer les jeux de données.

3.4 Exercices

Utilisation de la bonne loi. *Dans les exercices 1 à 4, effectuez l'une des actions appropriées : (a) trouver la valeur critique $z_{\alpha/2}$, (b) trouver la valeur critique $t_{\alpha/2}$, (c) établir que ni la loi normale ni la loi t ne s'applique.*

1. 95 % ; $n = 5$; σ est inconnu ; la population semble normale.

2. 99 % ; $n = 15$; σ est connu ; la population semble très asymétrique.

3. 90 % ; $n = 92$; σ est inconnu ; la population semble normale.

4. 98 % ; $n = 7$; $\sigma = 27$; la population semble normale.

Détermination des intervalles de confiance. *Dans l'exercice 5, utilisez le niveau de confiance et les données d'échantillon pour trouver (a) la marge d'erreur et (b) l'intervalle de confiance pour la moyenne μ de la population. On supposera que la population est normale.*

5. Test de maths/score pour les femmes : 95 %, $n = 15$, $\bar{x} = 496$, $s = 108$.

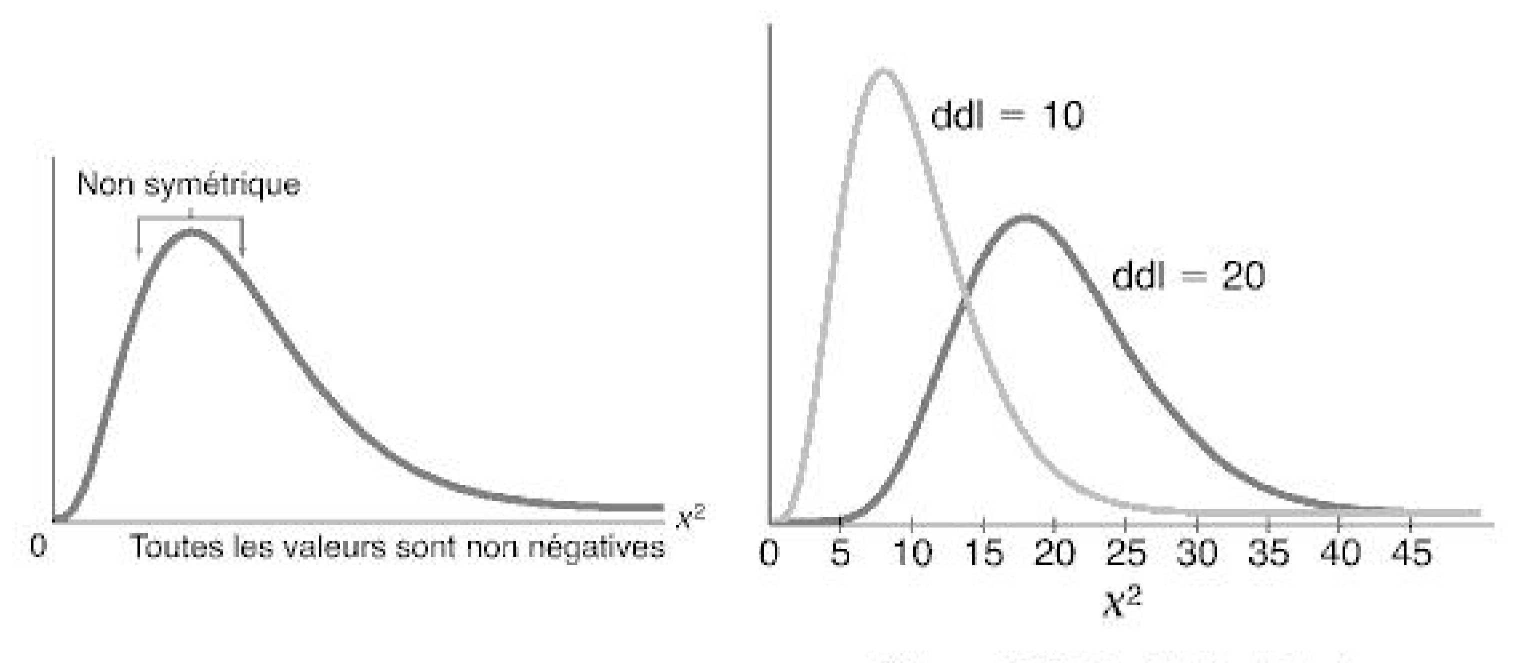


Figure 3.10 La loi du khi-deux

Figure 3.11 La loi du khi-deux pour ddl = 10 et ddl = 20

Exemple Valeurs critiques. Trouvez les valeurs critiques du χ^2 qui déterminent les régions critiques pour une aire de 0,025 dans chaque queue. On supposera que la taille associée de l'échantillon est 10 et donc que le nombre de degrés de liberté est 10 – 1 = 9.

Solution Regardez la figure 3.12 avec une table de khi-deux sous les yeux. La valeur critique à droite (χ^2 = 19,023) est obtenue directement en lisant la table à la ligne du nombre de degrés de liberté égal à 9 et à la colonne 0,025. La valeur critique à gauche (χ^2 = 2,700) correspond aussi à la ligne du nombre de degrés de liberté égal à 9 mais on utilise la colonne pour 0,975 = 1 – 0,025. La figure 3.12 montre que pour un échantillon de 10 valeurs prises dans une population normalement distribuée la statistique du khi-deux $(n - 1)s^2/\sigma^2$ a une probabilité 0,95 de tomber entre les valeurs critiques du khi-deux 2,700 et 19,023.

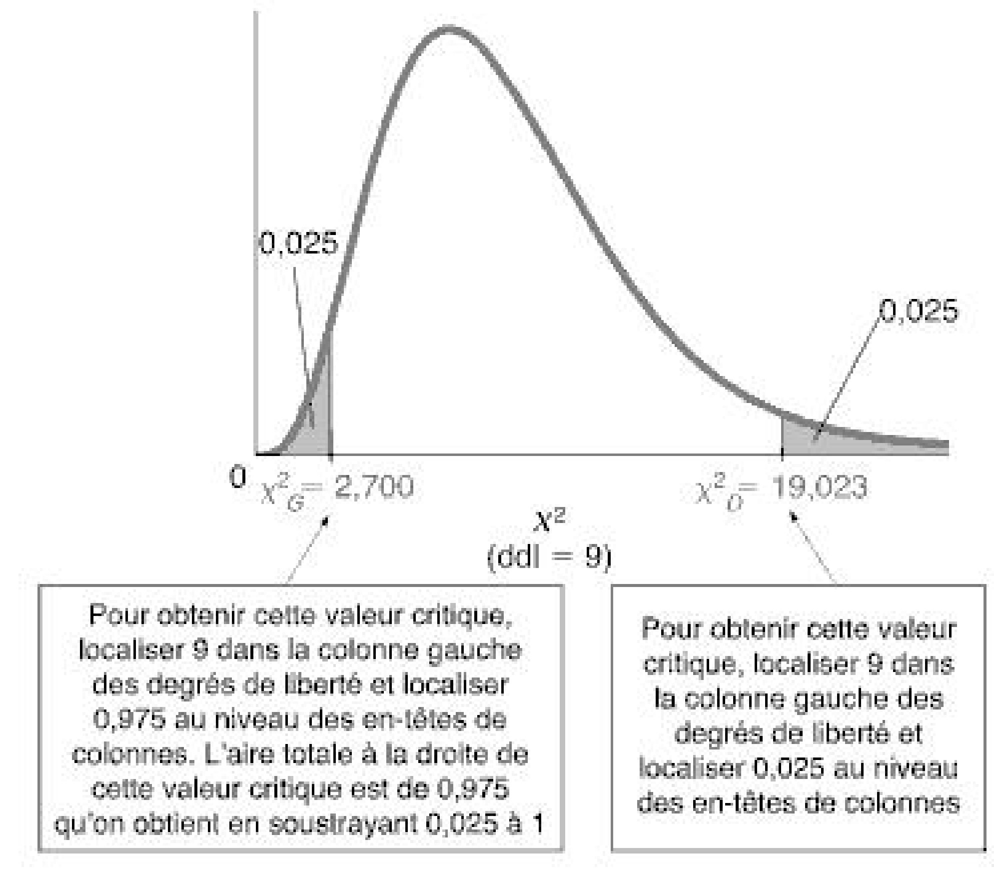
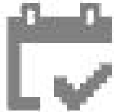


Figure 3.12 Valeurs critiques de la loi du khi-deux

Tableau 3-2 (suite)			
Taille d'échantillon pour σ^2		Taille d'échantillon pour σ	
Pour être sûr à 99 % que s^2 est à moins de	de la valeur de σ^2, la taille d'échantillon doit être au moins	Pour être sûr à 99 % que s est à moins de	de la valeur de σ, la taille d'échantillon doit être au moins
1 %	133 448	1 %	33 218
5 %	5 457	5 %	1 335
10 %	1 401	10 %	335
20 %	368	20 %	94
30 %	171	30 %	37
40 %	100	40 %	21
50 %	67	50 %	13



Exemple Nous voulons estimer l'écart type des températures corporelles. On veut être sûr à 95 % que notre estimation sera à moins de 10 % de la vraie valeur de σ. Quelle doit être la taille de l'échantillon ? On supposera la population normalement distribuée.

Solution À l'aide du tableau 3-2 on voit que le niveau de confiance à 95 % et une erreur de 10 % pour σ correspondent à une taille d'échantillon de 191. On devrait choisir aléatoirement 191 valeurs de la population des températures corporelles.

3.5 Exercices

Détermination des valeurs critiques. *Dans les exercices 1 et 2, trouvez χ^2_G et χ^2_D qui correspondent au niveau de confiance et à la taille d'échantillon donnés.*

1. 95 % ; $n = 16$.

2. 99 % ; $n = 80$.

Détermination des intervalles de confiance. *Dans les exercices 3 et 4, utilisez le niveau de confiance et les données d'échantillon pour trouver un intervalle de confiance de l'écart type σ de la population. Dans chaque cas, supposez qu'un échantillon aléatoire simple a été sélectionné d'une population normale.*

3. Salaires de professeurs de biologie : niveau de confiance 95 %, $n = 20$, $\bar{x} = 95\,000$ \$, $s = 12\,345$ \$.

4. Durées entre les utilisations de la télécommande du téléviseur par les hommes pendant la publicité : niveau de confiance 90 %, $n = 30$, $\bar{x} = 5{,}24$ s, $s = 2{,}50$ s.

Détermination des tailles d'échantillon. *Dans les exercices 5 et 6, supposez qu'un échantillon aléatoire simple a été sélectionné d'une population normale.*

5. Trouvez la taille d'échantillon minimale requise pour être sûr à 95 % que l'écart type s d'échantillon standard est à moins de 10 % de σ.

6. Trouvez la taille d'échantillon minimale requise pour être sûr à 99 % que la variance d'échantillon est à moins de 1 % de la variance de la population. Cette valeur est-elle une taille d'échantillon pratique pour la plupart des cas ?

Trouver des intervalles de confiance. *Dans les exercices 7 à 10, supposez que chaque échantillon est un échantillon aléatoire simple issu d'une population normale.*

7. ***Données historiques de maïs.*** Dans « The Probable Error of a Mean » de William Gosset (*Biometrika*, vol. VI, n° 1), paru en 1908, les valeurs suivantes étaient listées pour les rendements d'épis de maïs en kg par hectare. Ces valeurs correspondaient à des graines ordinaires (et non pas séchées au four). Construisez un intervalle de confiance à 95 % de l'écart type.

2 134 2 170 2 142 2 799 2 364 2 199 2 310 1 620 1 808 1 476 1 695

Utilisons le bon sens et pas de méthodes statistiques formelles. Que pourrions-nous conclure sur l'hypothèse de non-efficacité de « Choix du sexe » si 100 couples l'utilisent et ont 100 bébés avec :

a. 52 filles ?

b. 97 filles ?

Solution

a. On devrait normalement s'attendre à avoir environ 50 filles sur 100 naissances. Le résultat 52 est proche de 50 et on ne devrait pas conclure que « Choix du sexe » est efficace. Si les 100 couples n'utilisent aucune méthode spéciale pour choisir le sexe, le nombre 52 peut être dû au hasard. Avec 52 filles sur 100 naissances, il n'y a pas de preuves suffisantes pour dire que « Choix du sexe » est efficace.

b. Il est extrêmement improbable que le résultat 97 filles sur 100 naissances soit du au hasard. On peut expliquer la survenue de 97 filles de deux façons : soit un événement *extrêmement* rare est survenu par hasard, soit « Choix du sexe » est efficace. La probabilité extrêmement faible d'avoir 97 filles est une preuve forte contre l'hypothèse que « Choix du sexe » est inefficace. Dans ce cas, « Choix du sexe » semble être efficace.

Le point clé de l'exemple précédent est qu'on ne doit conclure à l'efficacité du produit que si on a de *façon significative* plus de filles qu'on ne devrait normalement en attendre. Bien que les résultats 52 filles et 97 filles sont tous deux « au-dessus de la moyenne », le résultat 52 filles n'est pas significatif alors que 97 filles est un résultat significatif.

Ce bref exemple illustre l'approche de base utilisée dans les tests d'hypothèse. La méthode formelle met en jeu un grand nombre de termes standard et de conditions intégrés dans une procédure organisée. Nous vous suggérons de commencer l'étude de ce chapitre en lisant d'abord rapidement les sections IV.2 et IV.3 pour vous faire une idée générale des concepts, puis de relire la section IV.2 avec plus d'attention pour devenir familier de la terminologie.

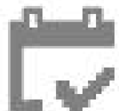
IV.2 Bases des tests d'hypothèse

Dans cette section, nous décrivons les composantes formelles utilisées dans les tests d'hypothèse : hypothèse nulle, hypothèse alternative, statistique de test, région critique, niveau de significativité, valeur critique, p-value, erreur de première espèce, de deuxième espèce. *On insiste dans cette section sur les composantes individuelles d'un test d'hypothèse alors que dans les sections suivantes nous considérerons ces composantes ensemble dans des procédures globales.*

Buts de cette section

- Étant donné une hypothèse, identifier l'hypothèse nulle, l'hypothèse alternative et les exprimer sous une forme symbolique.
- Étant donné une hypothèse et des données d'échantillon, calculer la valeur de la statistique de test.
- Étant donné un niveau de significativité, identifier la valeur critique.
- Étant donné la statistique de test, identifier la p-value.
- Établir la conclusion du test d'hypothèse en termes simples et non techniques.
- Identifier les erreurs de première et deuxième espèces qui pourraient être faites quand on teste une hypothèse donnée.

Vous devriez étudier l'exemple suivant jusqu'à ce que vous l'ayez compris en détail. Dès lors, vous aurez acquis un concept majeur des statistiques.



Exemple Identification de l'hypothèse nulle et de l'hypothèse alternative. À l'aide de la figure 4.2, utilisez les affirmations suivantes pour exprimer sous forme symbolique l'hypothèse nulle et l'hypothèse alternative correspondante :

a. La proportion de pois à gousse jaune est égale à 0,25.

b. La taille moyenne d'un homme adulte est au plus de 183 cm.

c. L'écart type pour la taille des femmes adultes est supérieur à 6 cm.

Solution

a. Pour l'étape 1 de la figure 4.2, on exprime l'affirmation donnée par $p = 0{,}25$. À l'étape 2, on voit que si $p = 0{,}25$ est fausse, alors $p \neq 0{,}25$ doit être vraie. Pour l'étape 3, on voit que dans les deux expressions $p = 0{,}25$ et $p \neq 0{,}25$ l'expression $p \neq 0{,}25$ ne contient pas d'égalité. On l'utilise donc comme hypothèse alternative. On pose donc $H_1 : p \neq 0{,}25$ et $H_0 : p = 0{,}25$.

b. Pour l'étape 1, on exprime « la moyenne est au plus de 183 cm » par la forme symbolique $\mu \leqslant 183$. À l'étape 2, on voit que si $\mu \leqslant 183$ est fausse, alors $\mu > 183$ doit être vraie. À l'étape 3, on voit que $\mu > 183$ ne contient pas d'égalité, donc on en fait l'hypothèse alternative. Donc $H_1 : \mu > 183$ et $H_0 : \mu = 183$.

c. Pour l'étape 1, on exprime l'affirmation sous la forme $\sigma > 6{,}0$. À l'étape 2, on voit que si $\sigma > 6{,}0$ alors $\sigma \leqslant 6{,}0$ doit être vraie. Pour l'étape 3, on prend comme hypothèse alternative $H_1 : \sigma > 6{,}0$ (parce qu'elle ne contient pas d'égalité) et donc H_0 est $\sigma = 6{,}0$.

Statistique de test

La statistique de test est une valeur calculée à partir des données d'échantillon et elle est utilisée dans la prise de décision du rejet ou non de l'hypothèse nulle. La statistique de test est obtenue par conversion de la statistique d'échantillon (comme la proportion d'échantillon $\hat{p}$, la moyenne d'échantillon $\bar{x}$ ou l'écart type d'échantillon s) en un score (comme z, t ou χ^2) en supposant que l'hypothèse nulle est vraie. La statistique de test peut alors être utilisée pour déterminer si on a suffisamment de preuves contre l'hypothèse nulle. Dans ce chapitre, nous considérons des tests d'hypothèse qui mettent en jeu des proportions, des moyennes ou des écarts types (ou des variances). Compte tenu des résultats des chapitres précédents sur les distributions d'échantillonnage pour les proportions, les moyennes et les écarts types, nous utiliserons les statistiques de test suivantes.

Statistique de test pour une proportion : $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$.

Statistique de test pour une moyenne : $z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ ou $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}}$.

Statistique de test pour un écart type : $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$.

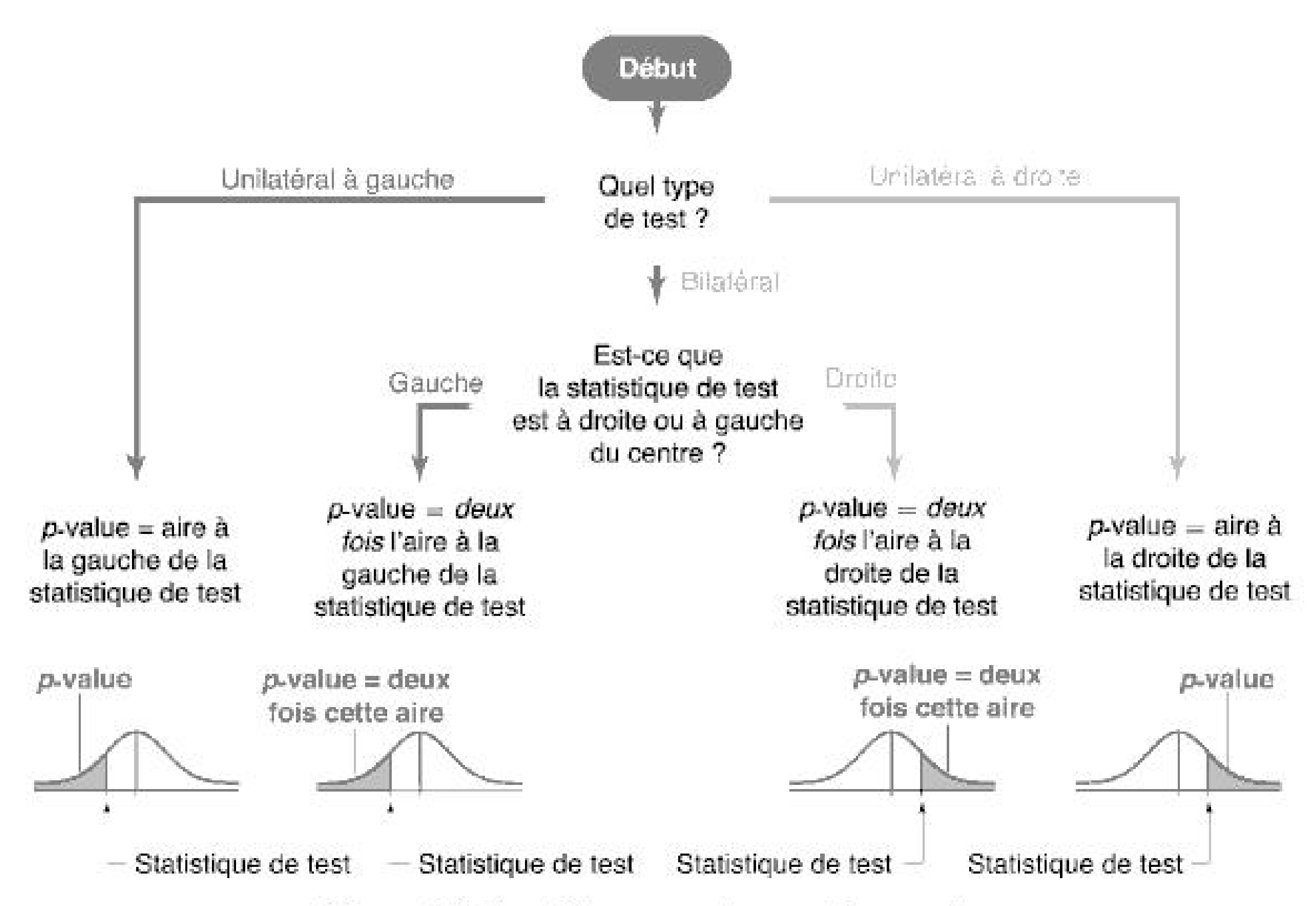


Figure 4.6 Procédure pour trouver les p-values

Décisions et conclusions

Nous avons vu que les affirmations originales deviennent parfois l'hypothèse nulle et d'autres fois l'hypothèse alternative. Cependant notre procédure standard pour tester une hypothèse oblige de toujours tester l'hypothèse nulle et donc notre conclusion initiale est toujours l'une des deux conclusions suivantes :

1. Rejet de l'hypothèse nulle.
2. Échec du rejet de l'hypothèse nulle.

Critère de décision. La décision de rejet ou d'échec de rejet pour l'hypothèse nulle est généralement faite soit selon la méthode traditionnelle (classique) de test d'hypothèse, soit selon la méthode de la p-value, soit encore en se basant sur les intervalles de confiance. Ces dernières années, l'usage de la méthode traditionnelle a décliné, en partie parce que les logiciels statistiques sont souvent conçus pour la méthode de la p-value.

Méthode traditionnelle : *Rejet* de H_0 si la statistique de test tombe dans la région critique.

Échec du rejet si la statistique de test ne tombe pas dans la région critique.

Méthode de la p-value : *Rejet* de H_0 si la p-value est $\leqslant \alpha$ (où α est le niveau de significativité, par exemple 0,05).

Échec du rejet si la p-value est $> \alpha$.

Autre option : Au lieu d'utiliser un niveau de significativité comme $\alpha = 0{,}05$ identifiez simplement la p-value et laissez la décision au lecteur.

Contrôler les erreurs de première et deuxième espèces : une étape dans notre procédure standard pour tester les hypothèses met en jeu la sélection du niveau de significativité α qui est la probabilité d'une erreur de première espèce. Par contre, on ne choisit pas β. Il serait souhaitable de toujours avoir $\alpha = 0$ et $\beta = 0$, mais dans la réalité ce n'est pas possible et on doit essayer de gérer les probabilités d'erreur α et β. Mathématiquement on peut montrer que α, β et la taille d'échantillon n sont liées et que quand on détermine deux des trois valeurs, la troisième est automatiquement fixée. Une pratique usuelle dans la recherche et dans l'industrie est de sélectionner α et n pour que β soit déterminé. Cependant une valeur de β supérieure à 0,2 est souvent considérée comme trop haute pour qu'un test d'hypothèse fournisse des résultats significatifs. Il faut essayer d'utiliser la plus grande valeur tolérable de α en fonction de l'impact de l'erreur de première espèce. Pour des erreurs de première espèce avec des conséquences très graves, il faut prendre des valeurs plus petites pour α. Prenez ensuite une taille d'échantillon n aussi grande que possible en tenant compte du coût, du temps et des autres facteurs pertinents. La détermination des tailles d'échantillon a été discutée dans les sections III.2 et III.3. Les considérations pratiques suivantes peuvent être utiles :

1. Pour α fixé, un accroissement de la taille d'échantillon n entraîne une diminution de β. C'est-à-dire qu'un plus grand échantillon diminuera la probabilité que vous commettiez l'erreur de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse.
2. Pour une taille d'échantillon n fixée, une diminution de α entraînera un accroissement de β. Réciproquement, un accroissement de α entraînera une diminution de β.
3. Pour diminuer à la fois α et β, il faut augmenter la taille d'échantillon n.

Pour donner du sens à ces idées abstraites, considérons les M&Ms (produits par la société Mars) et les plaquettes d'aspirine Bufferin (produites par la société Bristol-Meyers).

- Le poids moyen des bonbons M&Ms est supposé être d'au moins 0,9 g (de façon à être conforme au poids indiqué sur le paquet).
- Les plaquettes de Bufferin sont censées avoir un poids moyen de 325 mg d'aspirine.

Parce que les M&Ms sont des bonbons pour le plaisir alors que les plaquettes de Bufferin sont des médicaments pour traiter des problèmes de santé, on a affaire à deux niveaux très différents de gravité. Si les M&Ms n'ont pas un poids moyen de 0,9 g les conséquences ne seront pas très graves, mais si les plaquettes de Bufferin ne contiennent pas en moyenne 325 mg d'aspirine, les conséquences pourraient être importantes, ce qui inclut des actions en justice possibles et des poursuites de la part de l'Agence américaine de l'alimentation et des médicaments (FDA). En conséquence, pour tester l'affirmation que $\mu = 0{,}9$ g pour les M&Ms, on pourrait choisir $\alpha = 0{,}05$ et une taille d'échantillon $n = 100$; pour tester l'affirmation que $\mu = 325$ mg pour les plaquettes de Bufferin, on pourra choisir $\alpha = 0{,}01$ et une plus grande taille d'échantillon $n = 500$. Une plus grande taille d'échantillon nous permet de minimiser β tout en diminuant aussi α. Un plus petit niveau de significativité α et une plus grande taille d'échantillon n sont choisis à cause des conséquences plus graves associées au test d'un médicament public.

Puissance d'un test : on utilise β pour noter la probabilité de ne pas rejeter une hypothèse nulle fausse (erreur de deuxième espèce). Il s'ensuit que $1 - \beta$ est la probabilité de rejeter une hypothèse nulle fausse. Les statisticiens préfèrent nommer cette probabilité *puissance* du test et ils l'utilisent souvent pour juger l'efficacité d'un test à reconnaître qu'une hypothèse nulle est fausse. Une recommandation courante est de planifier une expérience pour que la puissance du test en résultant soit au moins de 0,8 (ou 80 %), de façon à ce que le test d'hypothèse soit très efficace à rejeter une hypothèse nulle fausse.

Identification de H_0 et H_1. *Dans les exercices 3 à 6, examinez les affirmations, puis exprimez l'hypothèse nulle H_0 et l'hypothèse alternative H_1 sous forme symbolique. Soyez sûrs d'utiliser le symbole correct (μ, p, σ) pour le paramètre indiqué.*

3. Le revenu moyen des travailleurs qui ont étudié les statistiques est supérieur à 50 000 $.

4. Plus de la moitié des utilisateurs d'Internet font des achats en ligne.

5. L'écart type de la taille des femmes est inférieur à 7,11 cm, qui est l'écart type de la taille des hommes.

6. La quantité moyenne d'alcool à 90 ° dans les conteneurs est au moins de 340 g.

Détermination des valeurs critiques. *Dans les exercices 7 à 10, trouvez les valeurs critiques z. Dans chacun des cas, supposez qu'on peut utiliser la loi normale.*

7. Test bilatéral $\alpha = 0{,}05$.

8. Test unilatéral $\alpha = 0{,}01$.

9. $\alpha = 0{,}10$; H_1 est $p \neq 0{,}19$.

10. $\alpha = 0{,}02$; H_1 est $p < 0{,}19$.

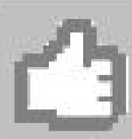
IV.3 Test d'hypothèse pour une proportion

Dans la section IV.2 nous avons présenté les composantes individuelles d'un test d'hypothèse mais dans cette section nous allons combiner ces composantes dans des tests d'hypothèse polyvalents pour des affirmations sur des proportions de population. Les proportions peuvent aussi représenter des probabilités ou l'équivalent décimal de pourcentages. Voici des exemples d'affirmations que nous serons capables de tester :

- Moins d'un quart des lycéens fument.
- Les sujets qui prennent le médicament Lipitor pour réduire leur cholestérol ont des maux de tête avec un taux supérieur à 7 %, qui est celui des gens qui ne prennent pas de Lipitor.
- À partir d'un sondage Gallup, la majorité (plus de 50 %) des Américains sont opposés au clonage humain.

Les suppositions requises, les notations et la statistique de test sont données ci-dessous. Les affirmations à propos d'une proportion sont habituellement testées à l'aide de la loi normale comme approximation de la loi binomiale.

Tester des hypothèses à propos d'une proportion p



1. Les observations d'échantillon proviennent d'un échantillon aléatoire simple. Il ne faut jamais oublier l'importance des méthodes d'échantillonnage.
2. Les conditions pour une loi binomiale sont satisfaites : il y a un nombre fixé d'essais indépendants avec une même probabilité et chaque essai n'a que deux résultats, nommés « succès » et « échec ».
3. Les conditions $np \geqslant 5$ et $nq \geqslant 5$ sont toutes les deux satisfaites ; ainsi **la loi binomiale peut être approximée par une loi normale avec $\mu = np$ et $\sigma = \sqrt{npq}$**.

Solution

Il faut d'abord vérifier que les conditions requises sont satisfaites. Compte tenu de l'expérience, il est raisonnable de supposer que l'échantillon est un échantillon aléatoire simple. Les conditions d'une expérience binomiale sont satisfaites parce qu'il y a un nombre fixe d'essais (77 + 420 018 = 420 095), que les essais sont indépendants (si un sujet a une leucémie, cela n'affecte pas la probabilité qu'un autre sujet l'ait), qu'il n'y a que deux résultats (leucémie ou pas) et que la probabilité de la leucémie reste constante. Au final, on utilise $n = 420\,095$ et $p = 0{,}000190$ pour voir que $np = 80 \geqslant 5$ et $nq = 420\,018 \geqslant 5$, donc la loi normale peut être utilisée pour approximer la loi binomiale. Les conditions sont satisfaites et le test peut être effectué. ❑

On utilise la méthode de la p-value de la figure 4.9. n vaut 420 095, $\hat{p} = 77/420\,095 = 0{,}000183$. Note : on pourrait s'arrêter ici car il n'y a aucune chance que la proportion d'échantillon $\hat{p} = 0{,}000183$ soit significativement plus grande que le taux supposé 0,000190. Mais nous continuerons pour l'exposé de la méthode.

Étape 1. L'affirmation originale est que le taux de leucémie pour les utilisateurs de téléphones portables est plus grand que 0,000190. Sous forme symbolique : $p > 0{,}000190$.

Étape 2. Son opposé est $p \leqslant 0{,}000190$.

Étape 3. Comme $p > 0{,}000190$ ne contient pas d'égalité, on en fait H_1. Soit :

H_0 : $p = 0{,}000190$ (hypothèse nulle).

H_1 : $p > 0{,}000190$ (hypothèse alternative et affirmation originale).

Étape 4. Le niveau de significativité est $\alpha = 0{,}01$.

Étape 5. Parce que l'affirmation met en jeu la proportion p, la statistique associée au test est la proportion d'échantillon $\hat{p}$ et la distribution d'échantillonnage est approximée par la loi normale parce que les conditions requises sont satisfaites.

Étape 6. La statistique de test vaut $z = -0{,}33$ trouvée comme suit :

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0{,}000183 - 0{,}000190}{\sqrt{\frac{0{,}000190 \times 0{,}999810}{420\,095}}} = -\,0{,}33.$$

Reportez-vous à la figure 4.6 pour trouver la p-value. Pour ce test unilatéral à droite, $z = -0{,}33$ a une aire de 0,3707 à gauche (lue dans une table de la loi normale), donc l'aire à droite est $1 - 0{,}3707 = 0{,}6293$. La p-value est donc 0,6293.

Étape 7. Parce que la p-value 0,6293 est plus grande que le niveau de significativité 0,01 on ne peut pas rejeter l'hypothèse nulle.

Interprétation On n'a pas pu rejeter l'hypothèse nulle, donc on considère qu'elle est correcte pour l'instant. On n'a pas pu confirmer l'hypothèse alternative qui était la conclusion originale. Voici la conclusion finale correcte : on n'a pas suffisamment de preuves pour garantir l'affirmation que les utilisateurs de téléphones portables ont un taux de leucémie supérieur à 0,000190 qui est celui des gens qui n'utilisent pas de téléphone portable. Il semble que les utilisateurs de téléphones portables ne risquent pas plus de développer une leucémie que les autres.

Méthode traditionnelle : à l'aide de la méthode traditionnelle on aurait les mêmes cinq premières étapes. À l'étape 6, la valeur critique serait $z = 2{,}33$. À l'étape 7, on ne pourrait pas rejeter l'hypothèse nulle parce que la statistique de test $z = -0{,}33$ ne tombe pas dans la région critique. Et on obtiendrait donc la même conclusion qu'avec la méthode de la p-value juste au-dessus.

Exemple Méthode de la p-value (voir figure 4.9)

Le jeu de données 2 de l'annexe B liste un échantillon de 106 températures humaines dont la moyenne est 36,78 °C. Supposez que l'échantillon est un échantillon aléatoire simple, que l'écart type σ est connu et vaut 0,34 °C. Utilisez un niveau de significativité de 0,05 pour tester la croyance commune que la température d'un adulte en bonne santé est 37,0 °C.

Solution

Adaptation Il faut d'abord vérifier que les conditions requises sont satisfaites, ce qui est le cas : reportez-vous à leur vérification dans l'exemple de la section III.3 car il s'agit des mêmes données. ❑

Les étapes suivantes correspondent à la figure 4.9 :

Étape 1. L'affirmation que la moyenne est 37,0 °C s'écrit symboliquement $\mu = 37{,}0$ °C.

Étape 2. L'opposé s'écrit symboliquement $\mu \neq 37{,}0$ °C.

Étape 3. Comme $\mu \neq 37{,}0$ °C ne contient pas d'égalité, on en fait l'hypothèse alternative, soit :

$H_0 : \mu = 37{,}0$ °C (affirmation originale)

$H_1 : \mu \neq 37{,}0$ °C.

Étape 4. Le niveau de significativité est spécifié dans l'énoncé : $\alpha = 0{,}05$.

Étape 5. Parce que l'affirmation est faite sur la moyenne μ de la population, la statistique de test la plus adaptée est la moyenne d'échantillon $\bar{x} = 36{,}78$ °C. Comme σ est supposé connu et que $n > 30$, le théorème de la limite centrale indique que la distribution des moyennes d'échantillon peut être approximée par la loi normale.

Étape 6. La statistique de test est calculée comme suit :

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{36{,}78 - 37{,}00}{\frac{0{,}34}{\sqrt{106}}} = -6{,}66$$

À l'aide de la statistique de test –6,66 on trouve la p-value associée qui doit être le double de l'aire à gauche de $z = -6{,}6$ car le test est bilatéral. À l'aide d'une table de la loi normale, l'aire à gauche est 0,0001 donc la p-value est 0,0002.

Étape 7. Comme la p-value 0,0002 est plus petite que le niveau de significativité $\alpha = 0{,}05$, on rejette l'hypothèse nulle.

Interprétation La p-value 0,0002 est la probabilité d'obtenir par hasard une moyenne d'échantillon aussi extrême que 36,78 °C (pour une taille d'échantillon $n = 106$) en supposant que $\mu = 37{,}0$ °C et $\sigma = 0{,}34$ °C. Comme cette probabilité est très petite, on rejette le hasard comme explication probable et on conclut que l'hypothèse $\mu = 37{,}0$ °C doit être fausse. À l'aide de la figure 4.7 de la section IV.2, on conclut qu'il y a suffisamment de preuves pour dire que la moyenne des températures diffère de 37,0 °C.

Exemple Températures. Une étudiante en médecine doit réaliser un projet de statistiques. Intriguée par les températures du jeu de données 2 de l'annexe B, elle décide de collecter ses propres données pour tester l'hypothèse que la moyenne des températures est inférieure à 37,0 °C. À cause de contraintes horaires imposées par les autres cours et son envie de maintenir une vie sociale (nocturne !) elle se rend compte qu'elle n'aura le temps de ne collecter que 12 valeurs. Après avoir planifié soigneusement la sélection d'un échantillon aléatoire simple de 12 adultes en bonne santé, elle obtient les températures listées ci-dessous. Utilisez un niveau de significativité de 0,05 pour tester l'hypothèse que la moyenne de ces températures est issue d'une population dont la moyenne est inférieure à 37,0 °C.

36,67	36,39	37,00	37,11	36,67	36,94
37,00	37,44	36,89	37,06	37,00	36,44

Solution

Il faut d'abord vérifier que les conditions requises sont satisfaites. On doit disposer d'un échantillon aléatoire simple, ce qui est le cas d'après l'énoncé. Ensuite $n = 12$ est inférieur à 30, donc il faut tester la normalité. L'histogramme suivant fourni par STATDISK (figure 4.12) montre que les données suivent une distribution pas très éloignée de la loi normale, donc le test peut être effectué. ❑

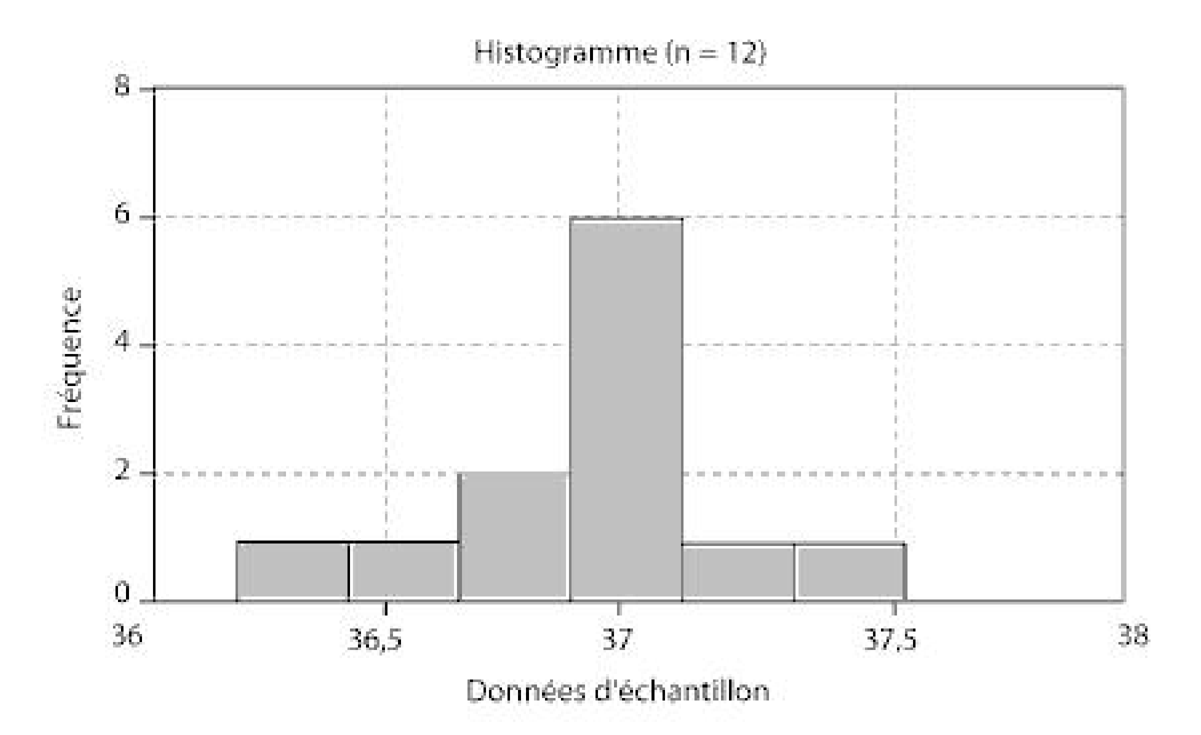


Figure 4.12

Après calcul sur les données, on a les statistiques d'échantillon suivantes : $\bar{x} = 36{,}88$ °C, $s = 0{,}297$ °C. La moyenne $\bar{x} = 36{,}88$ °C est inférieure à 37,0 °C, mais il faut déterminer si elle est significativement inférieure à 37,0 °C. On utilise les étapes de la figure 4.8.

Étape 1. L'affirmation originale que « la moyenne est inférieure à 37,0 °C » s'écrit symboliquement $\mu < 37{,}0$ °C.

Étape 2. L'opposé s'écrit symboliquement $\mu \geqslant 37{,}0$ °C.

Étape 3. Comme $\mu < 37{,}0$ °C ne contient pas d'égalité, on en fait l'hypothèse alternative, soit :

H_0 : $\mu = 37{,}0$ °C

H_1 : $\mu < 37{,}0$ °C (affirmation originale).

Étape 4. Le niveau de significativité est $\alpha = 0{,}05$.

Détermination des composantes. *Dans les exercices 5 et 6, supposez qu'un échantillon aléatoire simple a été sélectionné à partir d'une population normalement distribuée. Trouvez la statistique de test, la p-value, le(s) valeur(s) critique(s) et établissez la conclusion finale.*

5. Affirmation : la moyenne du score de QI pour les professeurs de statistiques est plus grande que 118. Données d'échantillon : $n = 20$, $\bar{x} = 120$, $s = 12$. Le niveau de significativité est $\alpha = 0,05$.

6. Affirmation : la durée moyenne entre les utilisations de la télécommande du téléviseur par les hommes pendant la publicité est égale à 5,00 s. Données d'échantillon : $n = 81$, $\bar{x} = 5,25$ s, $s = 2,50$ s. Le niveau de significativité est $\alpha = 0,01$.

Tests d'hypothèses. *Dans les exercices 7 à 10, supposez qu'un échantillon aléatoire simple a été choisi à partir d'une population normalement distribuée et testez l'affirmation donnée. Utilisez la méthode traditionnelle ou la méthode de la p-value.*

7. ***Effet d'un complément en vitamines sur le poids à la naissance.*** Des poids à la naissance (en kilogrammes) pour un échantillon d'enfants mâles nés de mères ayant eu un complément en vitamines ont été enregistrés (d'après des données du département de la santé de l'État de New York). Quand on teste l'affirmation que le poids moyen à la naissance pour de tels enfants est égal à 3,39 kg, qui est le poids moyen pour toute la population, SPSS affiche les résultats suivants. D'après ces résultats, est-ce qu'il semble qu'un complément en vitamines ait un effet sur le poids à la naissance ?

One-Sample Test

	Test Value = 3.39					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
WEIGHT	1.734	15	.103	.2849994	-.0652608	.6352595

Figure 4.14

8. ***Taille des parents.*** Le jeu de données 3 de l'annexe B inclut les tailles de 20 parents pour des enfants mâles. Si la différence de taille pour chaque couple de parents est obtenue en soustrayant la taille de la mère à celle du père, le résultat est une liste de 20 valeurs dont la moyenne est 11,18 cm et l'écart type 10,67 cm. Utilisez le niveau de significativité 0,01 pour tester l'affirmation que la différence moyenne est supérieure à 0. Est-ce que ces résultats confirment l'affirmation des sociologues que les femmes ont tendance à épouser des hommes plus grands qu'elles ?

9. ***Sucre dans les céréales.*** Un échantillon de paquets de céréales est sélectionné aléatoirement et le contenu en sucre (grammes de sucre par gramme de céréales) est enregistré. Ces quantités sont résumées par les statistiques : $n = 16$, $\bar{x} = 0,295$ g, $s = 0,168$ g. Utilisez le niveau de significativité 0,05 pour tester l'affirmation que le contenu moyen en sucre est inférieur à 0,3 g.

10. ***Gagnants aux Jeux Olympiques.*** On liste ci-dessous les temps gagnants (en secondes) des vainqueurs du 100 mètres hommes lors de jeux Olympiques d'été consécutifs classés par ordre décroissant en ligne. En supposant que ces résultats sont des données d'échantillon sélectionnées aléatoirement de la population de tous les jeux Olympiques passés et à venir, testez l'affirmation que la moyenne est inférieure à 10,5 s. Qu'observez-vous à propos de la précision de ces nombres ? Quelle caractéristique importante du jeu de données n'est pas prise en compte dans ce test d'hypothèse ? Est-ce que les résultats de ce test d'hypothèse suggèrent que les prochains temps gagnants seront autour de 10,5 s et est-ce qu'une telle conclusion est valide ?

12,0	11,0	11,0	11,2	10,8	10,8	10,8	10,6	10,8	10,3	10,3	10,3
10,4	10,5	10,2	10,0	9,95	10,14	10,06	10,25	9,99	9,92	9,96	

l'intervalle de confiance à 95 %, soit $5{,}2 < \sigma < 11{,}9$. Comme la valeur supposée $\sigma = 15$ n'est pas contenue dans l'intervalle de confiance, on rejette l'affirmation $\sigma = 15$ et on retrouve donc la même conclusion qu'avec les deux autres méthodes.

4.6 Exercices

Détermination des valeurs critiques. *Dans les exercices 1 et 2, trouvez la statistique de test, puis utilisez la table de la loi du χ^2 pour trouver le(s) valeur(s) critique(s) et les limites qui contiennent la p-value et, enfin, déterminez s'il y a suffisamment de preuves pour confirmer l'hypothèse alternative fournie.*

1. $H_1 : \sigma \neq 15$; $\alpha = 0{,}05$; $n = 20$; $s = 10$.

2. $H_1 : \sigma < 50$; $\alpha = 0{,}01$; $n = 30$; $s = 30$.

Test des affirmations à propos de la dispersion. *Dans les exercices 3 à 5, testez l'affirmation donnée. Supposez qu'un échantillon aléatoire simple a été sélectionné à partir d'une population normalement distribuée. On utilisera la méthode traditionnelle.*

3. ***Températures.*** Dans la section IV.4, nous avons testé l'affirmation que la température moyenne était égale à 37,0 °C et on a utilisé le jeu de données 2 de l'annexe B que l'on peut résumer par : $n = 106$, $\bar{x} = 36{,}78$ °C, $s = 0{,}34$ °C. Un histogramme montre que les valeurs ont une distribution approximativement normale. Dans la section IV.4, nous avons supposé que $\sigma = 0{,}34$ °C, ce qui est assez peu réaliste. Cependant la statistique de test causera le rejet de $\mu = 37{,}0$ °C tant que l'écart type sera inférieur à 1,17 °C. Utilisez les statistiques d'échantillon et un niveau de significativité 0,005 pour tester l'affirmation que $\sigma < 1{,}17$ °C.

4. ***Taille des top models.*** Utilisez un niveau de significativité 0,05 pour tester l'affirmation que la taille des femmes *top models* varie moins que celle des femmes en général. L'écart type de la taille des femmes est 6,35 cm. On liste ci-dessous la taille (en cm) de *top models* sélectionnées aléatoirement.

180	180	179	175	176	179	180	183	178
178	175	177	175	178	178	169	178	180

5. ***Est-ce que la nouvelle machine est meilleure ?*** La compagnie pharmaceutique Medassist utilise une machine pour verser des médicaments liquides dans des bouteilles de telle sorte que l'écart type du poids soit 4,25 g. Une nouvelle machine est testée sur 71 bouteilles et l'écart type pour cet échantillon est 3,4 g. La compagnie des machines Dayton qui fabrique la nouvelle machine affirme qu'elle remplit avec moins de dispersion. Testez au niveau de significativité 0,05 l'affirmation faite par la compagnie des machines Dayton. Si cette machine était utilisée en test, faudrait-il l'acheter ?

6. ***Poids des hommes.*** Des données d'une enquête anthropométrique sont utilisées pour publier des valeurs afin de permettre la fabrication de produits pour des adultes. Selon Gordon, Churchill *et al.*, les hommes ont un poids moyen de 78,1 kg avec un écart type de 13,1 kg. À l'aide de l'échantillon des poids des hommes listés dans le jeu de données 1 de l'annexe B, testez l'affirmation que l'écart type est de 13,1 kg. Utilisez le niveau de significativité 0,05. Quand on construit des ascenseurs, quelle serait la conséquence de croire que le poids des hommes varie moins qu'en réalité ?

7. ***Détermination des valeurs critiques de χ^2.*** Pour un nombre élevé de degrés de liberté, il est possible d'approximer les valeurs critiques du χ^2 comme suit :

$$\chi^2 = \frac{1}{2}\left(z + \sqrt{2k - 1}\right)^2$$

Ici k est le nombre de degrés de liberté et z est la valeur critique lue dans la table A-2. Par exemple, si on veut approximer les deux valeurs critiques du χ^2 dans un test d'hypothèse bilatéral avec $\alpha = 0{,}05$ et une taille d'échantillon de 150, on utilise $k = 149$ et $z = -1{,}96$ puis $k = 149$ et $z = +1{,}96$.

a. Utilisez cette approximation pour estimer les valeurs critiques du χ^2 dans un test d'hypothèse bilatéral avec $n = 101$ et $\alpha = 0{,}05$. Comparez ces résultats avec ceux trouvés dans la table A-4.

b. Utilisez cette approximation pour estimer les valeurs critiques du χ^2 dans un test d'hypothèse bilatéral avec $n = 150$ et $\alpha = 0{,}05$.

où $p_1 - p_2 = 0$ (supposé dans l'hypothèse nulle)

$$\widehat{p_1} = \frac{x_1}{n_1} \text{ et } \widehat{p_2} = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\bar{q} = 1 - \bar{p}.$$

p-value : utilisez la table A-2. Calculez la valeur de la statistique de test z et trouvez la p-value en suivant la procédure résumée à la figure 4.6.

Valeurs critiques : utilisez la table A-2. À partir du seuil de significativité α, trouvez les valeurs critiques en utilisant les procédures introduites dans la section III.2.

L'exemple suivant est utile pour clarifier les rôles de x_1, n_1, $\widehat{p_1}$, $\bar{p}$ et ainsi de suite. En particulier, on reconnaît que sous l'hypothèse de proportions égales, la meilleure estimation de la proportion commune est obtenue en combinant les deux échantillons ; ainsi $\bar{p}$ devient une estimation plus directe de la proportion de la population commune.

Exemple Test de l'efficacité d'un vaccin. Dans le problème introductif du chapitre, l'article de *USA Today* rapportait des résultats expérimentaux relatif à un vaccin administré à des enfants. Sur les 1 070 enfants ayant reçu le vaccin, 14 ont développé la grippe. Sur les 532 enfants qui ont reçu un placebo, 95 ont développé la grippe (tableau 5-1). Utilisez un seuil de significativité de 0,05 pour tester l'affirmation que la proportion d'enfants vaccinés qui développent la grippe est inférieure à celle des enfants qui ont reçu un placebo.

Solution On décide que l'échantillon 1 est le groupe recevant le traitement (vaccin) et que l'échantillon 2 est le groupe placebo. Nous pouvons résumer les données comme suit (les proportions $\widehat{p_1}$ et $\widehat{p_2}$ sont données avec des décimales supplémentaires car elles seront utilisées dans des calculs ultérieurs).

Enfants vaccinés	Enfants non vaccinés
$n_1 = 1\ 070$	$n_2 = 532$
$x_1 = 95$	$x_2 = 95$
$\widehat{p_1} = \frac{x_1}{n_1} = \frac{14}{1\ 074} = 0{,}013084$	$\widehat{p_2} = \frac{x_2}{n_2} = \frac{95}{532} = 0{,}178571$

Nous devons d'abord vérifier que les conditions requises sont satisfaites (voir plus haut). D'après le plan d'expérience, il est raisonnable de supposer que les deux échantillons sont des échantillons aléatoires simples et qu'ils sont indépendants. En outre, chaque échantillon comporte au moins 5 succès et 5 échecs : le premier contient 14 succès et 1 056 échecs, le second 95 succès et 437 échecs. Les conditions sont donc respectées et le test d'hypothèse formel peut être effectué. ❑

Nous allons utiliser la méthode de la p-value pour le test d'hypothèse, résumée à la figure 4.9.

Biostatistique

pour les sciences de la vie et de la santé

Édition revue et corrigée

La biostatistique est à la base de tout protocole de recherche scientifique, de la collecte des données à l'analyse et l'interprétation des résultats. Avec l'afflux de données génomiques et le développement croissant des études de populations, que ce soit en génétique, en médecine ou en écologie, la biostatistique n'est plus réservée à quelques spécialistes mais doit être maîtrisée par l'ensemble des biologistes et des cliniciens.

Le livre de Marc et Mario Triola se révèlera vite une **introduction à la biostatistique** indispensable :

- il ne requiert **aucune connaissance mathématique préalable** (le niveau terminale suffit)
- il est **complet, progressif et pédagogique**
- **il explique les concepts, détaille les formules et montre comment les appliquer** dans un souci constant de clarté et de simplicité
- les **nombreux graphiques** permettent de comprendre le cours en un coup d'œil
- **tous les exemples utilisent des données issues de la recherche en biologie** (ces données sont également disponibles sur le site web afin que vous puissiez reproduire les calculs)
- plus de **300 exercices corrigés** vous accompagnent dans la pratique de la biostatistique et l'évaluation de l'acquisition des concepts
- **des glossaires anglais/français et français/anglais et une liste des symboles statistiques** donnent accès aux principaux termes anglais employés dans les articles et les logiciels de biostatistique

Utilisable dès la première année d'université, cet ouvrage couvre tout le programme de biostatistique et vous accompagnera pendant toute la licence. Il constitue également une ressource précieuse pour toute personne désirant comprendre les fondements de la statistique, matière essentielle dans le monde d'aujourd'hui.

Public : étudiants en licence sciences de la vie et de la santé ou en classe préparatoire BCPST, écoles d'agronomie, STAPS ; peut aussi convenir aux étudiants de PACES et de DUT, BTS, IUT, IUP des domaines médicaux, pharmaceutiques et santé publique.

Niveau : Licence 1, 2, 3.

Marc M. Triola est médecin. Il dirige la section d'informatique médicale et le laboratoire des systèmes éducatifs avancés de l'École de médecine de l'université de New York. Il a totalement reprogrammé le logiciel de statistiques STATDISK®, et y a inclus des fonctions particulièrement utiles à la biostatistique.

Mario F. Triola est professeur émérite de mathématiques au Dutchess Community College, dans l'État de New York, où il a enseigné les statistiques pendant plus de 30 ans. Il est l'auteur de *Elementary Statistics*, maintenant à sa onzième édition. Il a rédigé plusieurs manuels et documents de travail pour la technologie dédiée à l'enseignement des statistiques.

Traduction française :
Gilles Hunault et Yves Desdevises

Web ressources

Le site www.compagnons.pearson.fr/biostatistique (en anglais) contient les jeux de données utilisés dans divers formats (Excel®, SAS®, SPPS®, JMP®,...), le logiciel de statistiques STATDISK® et le programme DDXL® qui ajoute un menu supplémentaire dans Excel et permet ainsi de calculer des macros statistiques.

ISBN : 978-2-7440-7657-2

7657 1112 42,50 €



www.pearson.fr

ALWAYS LEARNING

PEARSON